

# Projet en atelier statistique avec R

## Analyse de l'Impact des Performances des Joueurs sur les Résultats d'équipe

Yassine Hammi

2024-11-26

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Question principale d'exploration . . . . .	2
1.2	Spécification des Variables . . . . .	2
<b>2</b>	<b>Collecte de Données</b>	<b>2</b>
2.1	Identification . . . . .	2
2.2	Importation . . . . .	2
<b>3</b>	<b>PreProcessing</b>	<b>4</b>
3.1	Suppression des colonnes dans la première dataset : RealMadridPlayers . . . . .	4
3.2	Suppression des lignes redondantes dans la première dataset : RealMadridPlayers . . . . .	5
3.3	Regroupement des joueurs par leur positions de jeu . . . . .	6
3.4	Supprimer les colonnes non nécessaires pour chaque groupe . . . . .	8
3.5	Nettoyage du dataset des matches RM . . . . .	10

## 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 individuels 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 permettront 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 performances pour chaque joueur individuelle et comme conséquences la performance d'équipe et comment cela influence l'échec ou la réussite de l'équipe.

## 1.2 Spécification des Variables

Pour mener à bien notre analyse, nous nous appuyerons sur différentes sources de données cruciales telles que "Performance des joueurs individuels" et leur "Résultats 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 structurées, 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 distinctes.

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

# Chargement de package
library(readr)
library(knitr)

# Importation de données depuis fichier excel
# Cette fichier est importé depuis kaggle
# Comporte les données 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 données
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
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 trouvé")
}
```

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 supprimé 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 ) )

# 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[,1:9],6 ),
  caption = "Tableau des Données 1er 6 lignes avec 9 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 = 8
  )

#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
```

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

Table 3: Tableau des Données 1er 6 lignes avec 9 colonnes

name	shirt_number	position	aerial_duels	aerial_duels_lost	aerial_duels_won	appearances	assists_intentional	attempts_from_set_pieces
Andrii Lunin	13	Goalkeeper	10	NA	10	21	NA	NA
Antonio Rüdiger	22	Defender	72	23	49	33	NA	NA
Arda Güler	24	Midfielder	NA	NA	NA	10	NA	1
Aurélien Tchouaméni	18	Midfielder	75	24	51	27	1	NA
Brahim Díaz	21	Midfielder	12	9	3	31	5	NA
Dani Carvajal	2	Defender	41	19	22	28	3	NA

### 3.2 Suppression les lignes redondantes dans la première dataset : RealMadridPlayers

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

# Identifier les lignes dupliquées.

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>, ...
```

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

kable(head(normaldt_RM_Players[,1:9],6),align = "l")
```

name	shirt_number	position	aerial_duels	aerial_duels_lost	aerial_duels_won	appearances	assists_intentional	attempts_from_set_pieces
Andrii Lunin	13	Goalkeeper	10	NA	10	21	NA	NA
Antonio Rüdiger	22	Defender	72	23	49	33	NA	NA
Arda Güler	24	Midfielder	NA	NA	NA	10	NA	1
Aurélien Tchouaméni	18	Midfielder	75	24	51	27	1	NA
Brahim Díaz	21	Midfielder	12	9	3	31	5	NA
Dani Carvajal	2	Defender	41	19	22	28	3	NA

### 3.3 Regroupement les joueurs 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

```
#
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)
```

Table 7: Midfielders Table

name	shirt_number	position	aerial_duels	aerial_duels_lost
Arda Güler	24	Midfielder	NA	NA
Aurélien Tchouaméni	18	Midfielder	75	24
Brahim Díaz	21	Midfielder	12	9
Dani Ceballos	19	Midfielder	7	5
Eduardo Camavinga	12	Midfielder	36	18
Federico Valverde	15	Midfielder	31	10

```
#
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)
```

Table 8: Forwards Table

name	shirt_number	position	aerial_duels	aerial_duels_lost
Gonzalo García	33	Forward	NA	NA
José Luis Mato Sanmartín	14	Forward	77	36
Lucas Vázquez	17	Forward	10	6
Rodrygo Silva de Goes	11	Forward	26	21

Vinícius José Paixão de Oliveira Júnior	7	Forward	8	7
Álvaro Rodríguez	29	Forward	NA	NA

---

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

### 3.4 Supprimer les colonnes non nécessaires pour chaque groupe

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\_distribution, 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 colonnes 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)
}

GoalKeepers <- remove_na_columns(GoalKeepers)

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

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

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

```
## [1] "Dimension Avant suppression"
## [1] 6 59
## [1] "Dimension Après suppression"
## [1] 6 33
```



Répéter le processus pour toutes les positions :

- pour défenseurs:

```
## [1] "Dimension Avant suppression"
## [1] 11 110
## [1] "Dimension Après 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 Après suppression"
## [1] 11 55
```

- 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 Après 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 Après suppression"
## [1] 12 48
```

- pour les attaquants:

```
## [1] "Dimension Avant suppression"
## [1] 6 110
## [1] "Dimension Après 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 Après suppression"
## [1] 6 28
```

### 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 existing 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")

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

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

delete_columns<- function(MatchesRM, columnsToDelete) {
  # si existe column -> put it in existing 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")

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