Rapport

Ahamad MOHAMMAD; Minko Bikono NEIL-JOVY; Simon GELBART; Willen AMICHE

2025-04-27

Introduction

Notre objectif est d'explorer l'impact des stratégies de recrutement sur les performances sportives des clubs et joueurs. En combinant des statistiques individuelles, collectives et des données de transferts, nous chercherons à identifier les tendances qui influencent la réussite des équipes sur plusieurs saisons.

Description des variables

- **Player Stats 2021-2022** (`2021-2022 Football Player Stats.csv`) 143 variables – chaque ligne correspond à un joueur pour la saison 2021-2022.

Variable	Type	Description approximative
`Rk`	int64	Rang ou ID du joueur
`Player`	object	Nom du joueur
`Nation`	object	Nationalité
`Pos`	object	Poste
`Squad`	object	Club
`Comp`	object	Compétition principale
`Age`	object	Âge
`Born`	object	Année de naissance
…	…	…et plus de 130 autres statistiques de jeu : buts, passes, tirs, dribbles, fautes, tacles, interceptions, passes progressives, etc.
`AerWon`	float64	Duels aériens gagnés
`AerLost`	float64	Duels aériens perdus
`AerWon%`	float64	Pourcentage de duels aériens gagnés

- **Team Stats 2021-2022** (`2021-2022 Football Team Stats.csv`) 20 variables – chaque ligne correspond à une équipe.

Variable	Type	Description
`Rk`	int64	Rang
`Squad`	object	Nom du club
`Country`	object	Pays
`LgRk`	int64	Classement dans la ligue

Variable	Type	Description
MP, W, D, L	int64	Matchs joués, Victoires, Nuls, Défaites
GF, GA, GD	int64	Buts pour, contre, différence
Pts, Pts/G	int64 / float64	Points et moyenne par match
xG, xGA, xGD, xGD/90	float64	Données d'expected goals
Attendance	int64	Affluence moyenne
Top Team Scorer	object	Meilleur buteur
Goalkeeper	object	Gardien principal

- **Team Stats 2022-2023** (`2022-2023 Football Team Stats.csv`) Même structure et signification que pour 2021-2022, mais avec la saison suivante.

- **Transfers Été 2022** (`2022_2023_football_summer_transfers.csv`) 11 variables – chaque ligne correspond à un transfert.

Variable	Type	Description
name	object	Nom du joueur transféré
position	object	Poste
age	object	Âge
market_value	object	Valeur estimée
country_from	object	Pays de départ
league_from	object	Ligue de départ
club_from	object	Club de départ
country_to	object	Pays d'arrivée
league_to	object	Ligue d'arrivée
club_to	object	Club d'arrivée
fee	object	Montant du transfert (peut contenir "Free", "Loan", etc.)

Analyse et réponses aux questions

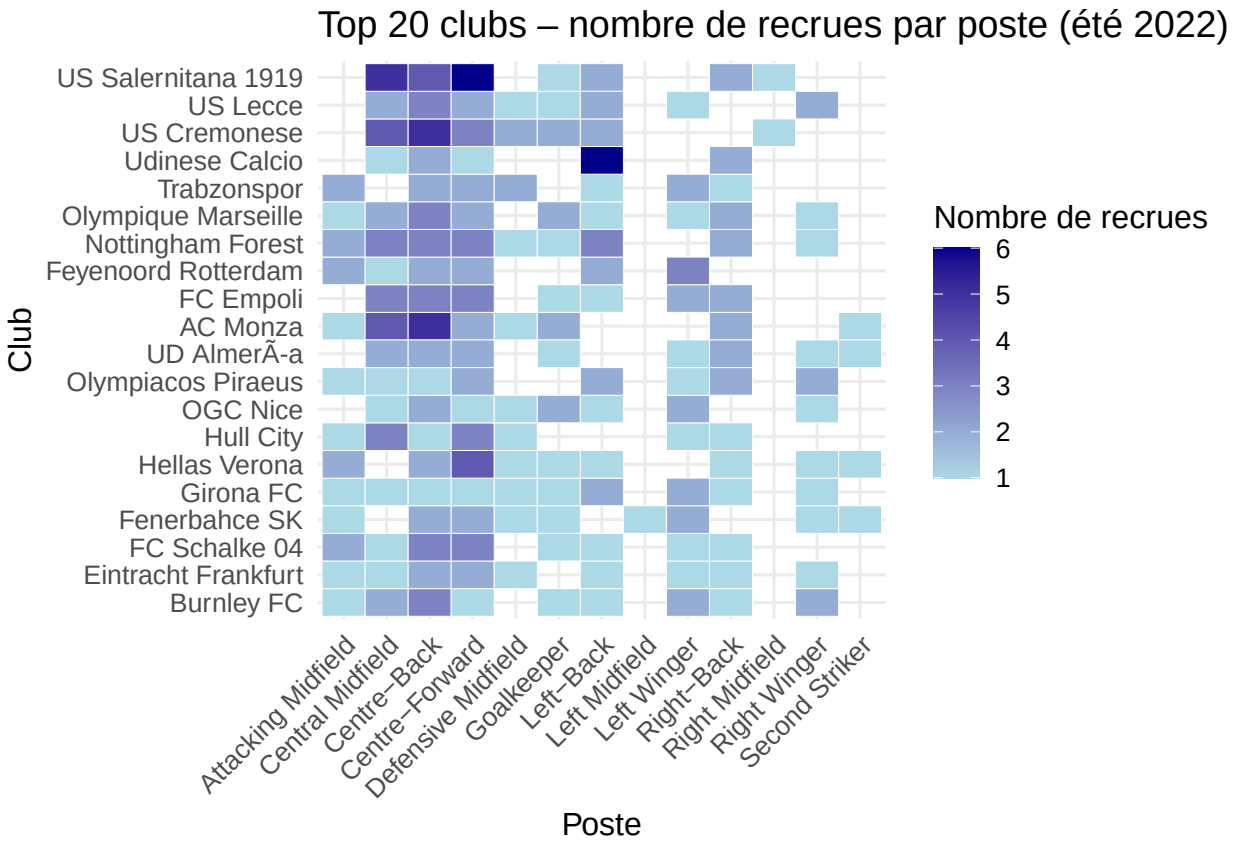
```r
# Chargement des packages nécessaires

library(tidyverse)
library(ggplot2)
```

```r
# Importation des données

player_stats_2021_2022 <- read.csv("data/2021-2022 Football Player Stats.csv", sep = ";", fileEncoding =
team_stats_2021_2022 <- read.csv("data/2021-2022 Football Team Stats.csv", sep = ";", fileEncoding = "I
team_stats_2022_2023 <- read.csv("data/2022-2023 Football Team Stats.csv", sep = ";", fileEncoding = "I
transfers_2022 <- read.csv("data/2022_2023_football_summer_transfers.csv", sep = ",", fileEncoding = "I
```

Question 1: Quels clubs ont le plus recruté par poste (top 20) ?

Objectif: Identifier les **20 clubs les plus actifs** lors du mercato d'été 2022 en termes de nombre de recrues, puis visualiser quels postes ont été ciblés en priorité par ces clubs. Cela permet de mieux comprendre les **stratégies de renforcement** des effectifs selon les lignes de jeu (défense, milieu, attaque).

```r
# Code

# Nettoyage de base
transfers_clean <- transfers_2022 %>%
  filter(!is.na(club_to), !is.na(position), position != "")

# Regrouper et compter
recrutements_par_poste <- transfers_clean %>%
  group_by(club_to, position) %>%
  summarise(nb_recrues = n(), .groups = "drop")

# Garder uniquement les 20 clubs ayant recruté le plus globalement
top_clubs <- recrutements_par_poste %>%
  group_by(club_to) %>%
  summarise(total_recrues = sum(nb_recrues)) %>%
  top_n(20, total_recrues) %>%
  pull(club_to)

# Filtrer les données
recrutements_top <- recrutements_par_poste %>%
  filter(club_to %in% top_clubs)

# Visualisation
ggplot(recrutements_top, aes(x = position, y = fct_reorder(club_to, nb_recrues), fill = nb_recrues)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(
    title = "Top 20 clubs - nombre de recrues par poste (été 2022)",
    x = "Poste",
    y = "Club",
    fill = "Nombre de recrues"
  ) +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interprétation du graphique:

- **US Salernitana 1919** est le club qui a le plus recruté tous postes confondus, avec un **focus important sur les milieux offensifs** et **défenseurs centraux**.
- **US Lecce**, **US Cremonese** et **Udinese Calcio** (clubs italiens) montrent aussi une stratégie de **renforcement défensif**, particulièrement en **centre-back**.
- **Olympique de Marseille** et **Nottingham Forest** ont **diversifié leurs recrutements** sur plusieurs lignes, y compris **les ailes** (*left/right winger*).
- On observe une **forte demande en milieux de terrain**, notamment :
  - *Attacking Midfield*
  - *Defensive Midfield*
  - *Centre Midfield*
- Peu de clubs ont recruté plusieurs **gardiens**, ce qui est logique : un club n'en fait souvent venir qu'un seul par saison.
- Certains clubs comme **FC Empoli** ou **OGC Nice** présentent une stratégie de recrutement **équilibrée sur différentes lignes**, ce qui pourrait indiquer un **renouvellement global de l'effectif**.

Question 2: Quels postes sont les plus valorisés sur le marché ?

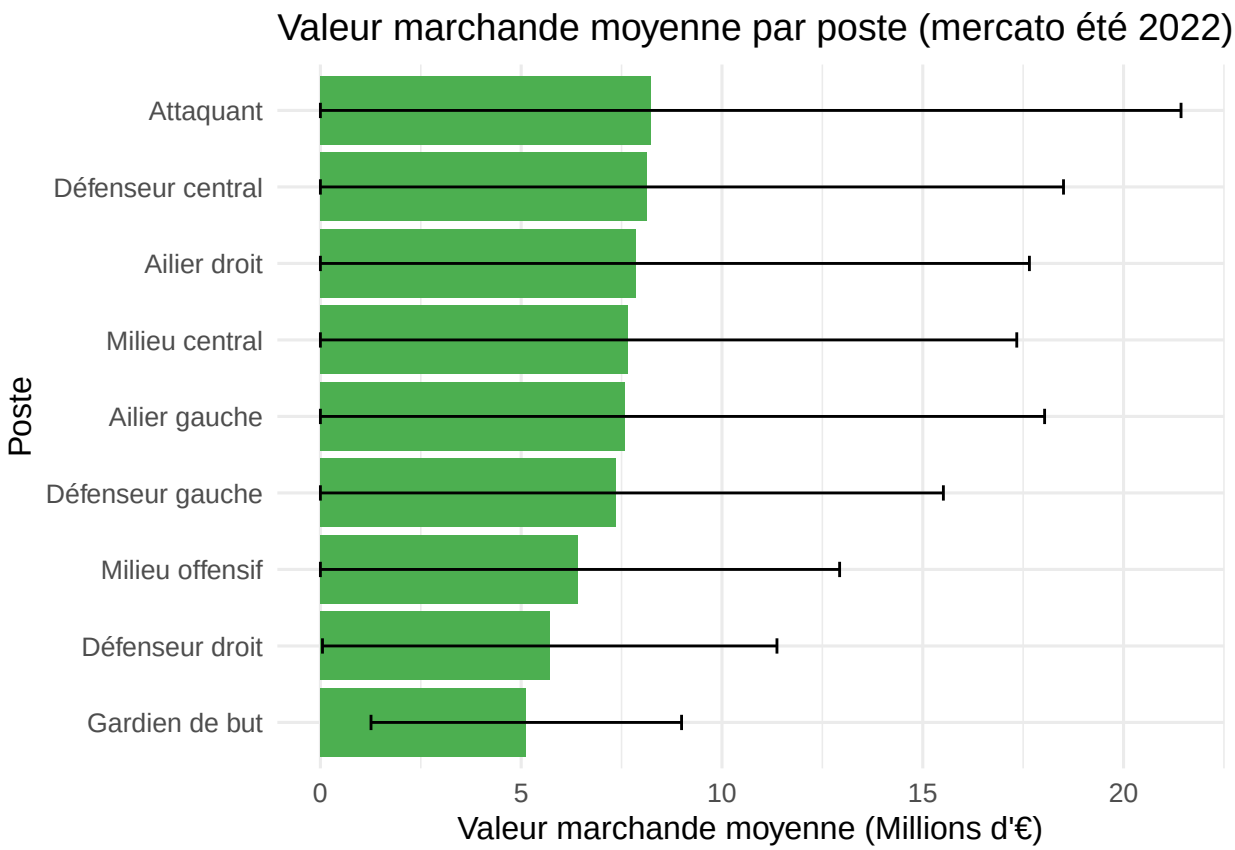
Objectif: Visualiser la valeur marchande des joueurs par poste pour comprendre quels types de profils sont les plus prisés financièrement sur le marché. Cela permet de hiérarchiser les postes selon leur importance économique dans le football professionnel. Pour ça, on se base sur le mercato qui a eu lieu durant l'été 2022.

```
# Nettoyage des données
transfers_clean <- transfers_2022 %>%
  filter(!is.na(position), !is.na(market_value)) %>%
  distinct(name, position, age, .keep_all = TRUE) %>%
  mutate(
    market_value_num = as.numeric(market_value)  # Convertir les valeurs en numérique
  ) %>%
  filter(market_value_num >= 1.5)  # Filtrer les valeurs négatives/nulles ou pas réalistes genre en des
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `market_value_num = as.numeric(market_value)`.
## Caused by warning:
## ! NAs introduits lors de la conversion automatique
```

```
# On crée une colonne pour regrouper certains postes qui sont très proches
transfers_clean <- transfers_clean %>%
  mutate(position_fr = position) %>%
  mutate(position_fr = recode(position_fr,
                              "Second Striker" = "Attaquant",
                              "Centre-Forward" = "Attaquant",
                              "Forward Attacker" = "Attaquant",
                              "attack" = "Attaquant",
                              "Right Midfield" = "Ailier droit",
                              "Left Midfield" = "Ailier gauche",
                              "Attacking Midfield" = "Milieu offensif",
                              "Central Midfield" = "Milieu central",
                              "Defensive Midfield" = "Milieu central",
                              "Centre-Back" = "Défenseur central",
                              "defence" = "Défenseur central",
                              "Right Winger" = "Ailier droit",
                              "Left Winger" = "Ailier gauche",
                              "Left-Back" = "Défenseur gauche",
                              "Right-Back" = "Défenseur droit",
                              "Goalkeeper" = "Gardien de but",
                              .default = position_fr))  # Par défaut, on garde la position d'origine

# Calcul de la moyenne et de l'écart-type des valeurs marchandes par poste
transfers_clean_stats <- transfers_clean %>%
  group_by(position_fr) %>%
  filter(n() >= 5) %>%
  summarise(
    mean_value = mean(market_value_num, na.rm = TRUE),
    sd_value = sd(market_value_num, na.rm = TRUE)
  )

# Visualisation avec un barplot et des barres d'erreur pour voir les postes qui ont une grosse différen
ggplot(transfers_clean_stats, aes(x = reorder(position_fr, mean_value), y = mean_value)) +
```

```
  geom_bar(stat = "identity", fill = "#4CAF50") +
  geom_errorbar(aes(ymin = pmax(mean_value - sd_value, 0), ymax = mean_value + sd_value), width = 0.2) +
  labs(
    title = "Valeur marchande moyenne par poste (mercato été 2022)",
    x = "Poste",
    y = "Valeur marchande moyenne (Millions d'€)"
  ) +
  theme_minimal(base_size = 12) +
  coord_flip()
```




```
transfers_clean %>%
  count(position_fr, sort = TRUE)
```

```
##         position_fr   n
## 1    Milieu central 271
## 2         Attaquant 245
## 3 Défenseur central 203
## 4     Ailier gauche 121
## 5      Ailier droit  96
## 6   Milieu offensif  95
## 7  Défenseur gauche  82
## 8   Défenseur droit  76
## 9    Gardien de but  63
```

```r
print(transfers_clean %>% count(position_fr, sort = TRUE))
```

```
##         position_fr   n
## 1    Milieu central 271
## 2         Attaquant 245
## 3 Défenseur central 203
## 4     Ailier gauche 121
## 5      Ailier droit  96
## 6   Milieu offensif  95
## 7  Défenseur gauche  82
## 8   Défenseur droit  76
## 9    Gardien de but  63
```

Interprétation du graphique: Attaquant et Buteur : Les attaquants sont les joueurs les plus valorisés sur le marché, avec une valeur moyenne qui dépasse largement celle des autres postes. Cela reflète l'importance des joueurs offensifs dans le football moderne, où les buts et les performances offensives sont souvent décisives mais surtout les attaquants sont souvent les joueurs qui attirent le plus l'attention des fans et des médias, ce qui en fait des atouts marketing précieux pour les clubs et qui explique leur valeur bien plus élevé

Défenseur central : Les défenseurs centraux occupent la deuxième place en termes de valeur marchande. Leur rôle crucial dans la stabilité défensive et leur capacité à organiser la défense expliquent cette valorisation élevée. Un bon défenseur central peut être la clé de voûte d'une équipe solide et équilibrée.

Ailier droit et Ailier gauche : Les ailiers, qu'ils soient droits ou gauches, présentent une valeur marchande significative sur le marché des transferts avec respectivement la 3ème et 5ème place. Toutefois, on observe souvent une légère survalorisation des ailiers droits. Cette différence peut s'expliquer par la rareté relative des ailiers droits car cela implique souvent d'être gaucher, pour être capables de repiquer dans l'axe et de frapper avec leur pied fort. Ce profil est très recherché dans le football moderne d'où une valorisation plus élevée.

Milieu central et offensif : Les milieux centraux, souvent considérés comme les "poumons" de l'équipe, sont également très valorisés. Leur polyvalence et leur capacité à influencer le jeu dans les deux sens en font des acteurs clés sur le marché. Paradoxalement, les milieux offensifs semblent être moins bien valorisés. C'est un poste qui, dans le football moderne, a perdu de sa splendeur par rapport à avant et c'est caractérisé par une baisse de la valeur marchande à ce poste.

Défenseur gauche et Défenseur droit : Les défenseurs latéraux, bien que moins valorisés que les défenseurs centraux, restent importants.Ils sont très recherchés dans le football moderne car ils doivent être polyvalents, capables de défendre et d'attaquer efficacement.On ne retrouve pas la rareté des latéraux droit par rapport au gauche comme on peut avoir avec les ailiers et on remarque qu'à l'inverse les latéraux gauche on tendance à avoir une plus grosse valeur marchande.

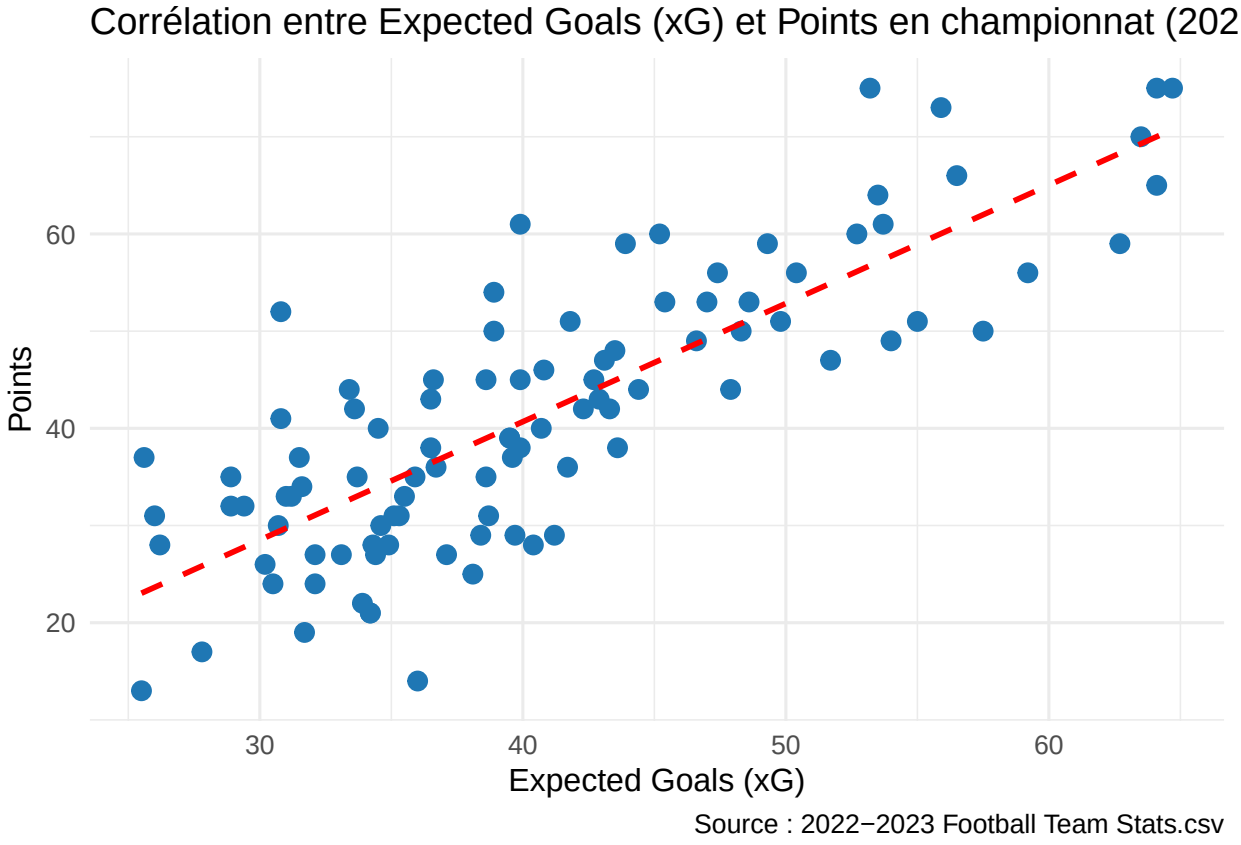
Gardien de but : Les gardiens de but sont nettement moins valorisés que les autres postes. Cela s'explique en partie par la spécificité de ce rôle, qui diffère fortement des autres positions sur le terrain. Par extension, c'est aussi le poste le moins "vendeur" du football, ce qui se reflète dans leur valeur marchande.

Question 3: Corrélation entre xG et points ?

Objectif: Analyser si le nombre d'expected goals (xG) réalisés par une équipe est corrélé avec son total de points en championnat. Cela permet de vérifier si produire beaucoup d'occasions (même sans forcément marquer) est un bon indicateur de performance globale. On utilise les données de la saison 2022-2023.

```r
# On utilise les données de la saison 2022-2023
team_stats_22_23 <- team_stats_2022_2023 %>%
  select(Squad, xG, Pts)

# Nettoyage basique : on supprime les NA au cas où
team_stats_22_23 <- team_stats_22_23 %>%
  filter(!is.na(xG), !is.na(Pts))

# Visualisation scatter plot
ggplot(team_stats_22_23, aes(x = xG, y = Pts)) +
  geom_point(color = "#1f77b4", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +
  labs(
    title = "Corrélation entre Expected Goals (xG) et Points en championnat (2022-23)",
    x = "Expected Goals (xG)",
    y = "Points",
    caption = "Source : 2022-2023 Football Team Stats.csv"
  ) +
  theme_minimal(base_size = 12)
```

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




```r
# Calcul du coefficient de corrélation de Pearson
cor_xg_pts <- cor(team_stats_22_23$xG, team_stats_22_23$Pts, use = "complete.obs")
paste0("Coefficient de corrélation de Pearson : ", round(cor_xg_pts, 3))
```

```
## [1] "Coefficient de corrélation de Pearson : 0.815"
```

Interprétation du graphique: Le graphique montre qu'il existe une forte corrélation positive entre les expected goals (xG) et les points obtenus en championnat lors de la saison 2022-2023. Le coefficient de corrélation de Pearson est de 0,815, ce qui indique qu'en général, plus une équipe génère d'occasions de but de qualité (mesurées par les xG), plus elle obtient de points au classement. Cela confirme que produire beaucoup d'occasions est un élément important pour réussir sur toute une saison.

Cependant, la relation n'est pas parfaite. On observe que certains points sont assez éloignés de la tendance générale. Cela peut s'expliquer par plusieurs facteurs, comme l'efficacité offensive, c'est-à-dire la capacité à transformer les occasions en buts, ou encore la solidité défensive. Une équipe peut créer beaucoup d'occasions mais manquer de réalisme ou encaisser trop de buts, ce qui limite son total de points.

En résumé, les expected goals sont un bon indicateur des performances d'une équipe, mais ils ne suffisent pas à eux seuls pour expliquer complètement les résultats. D'autres aspects du jeu, comme la finition ou la défense, restent essentiels.

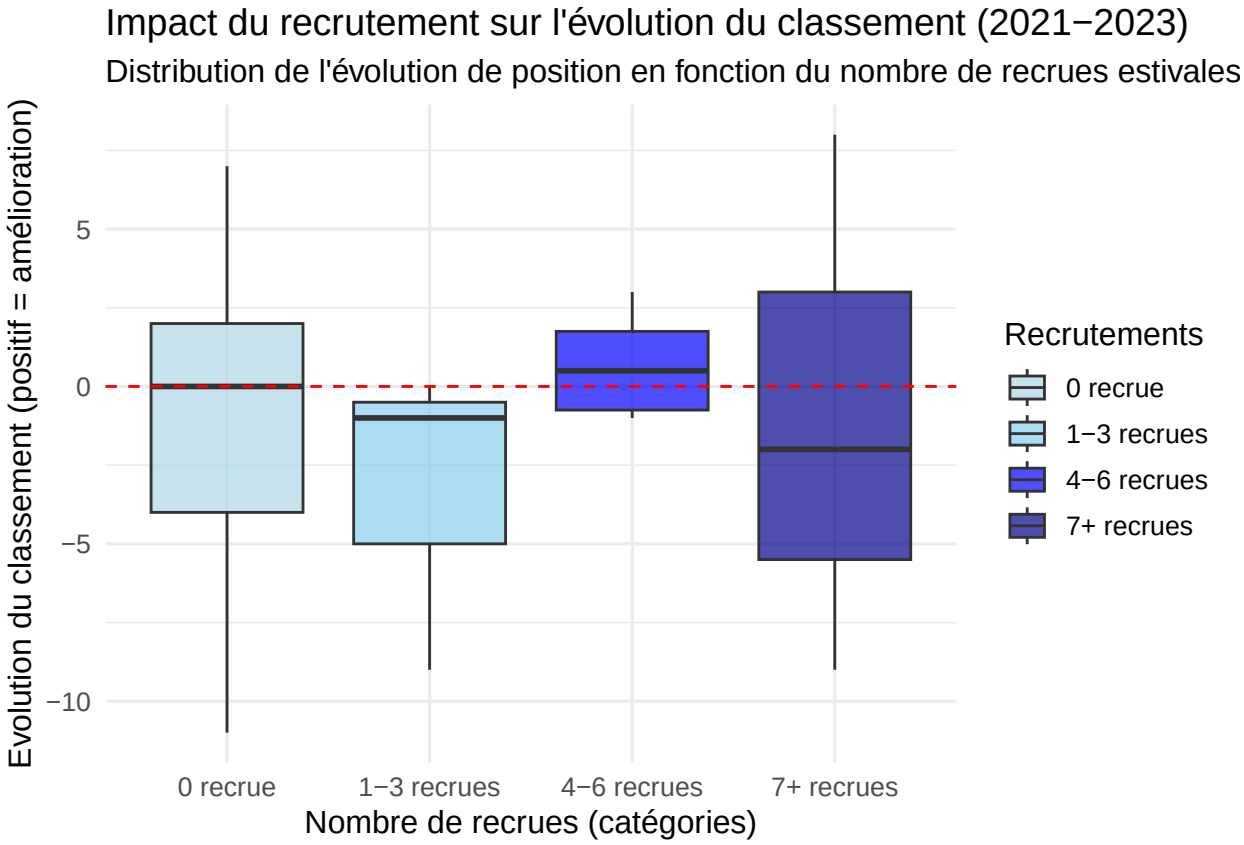
Question 4: Quel est l'impact des transferts sur la performance des clubs entre 2021-2022 et 2022-2023 ?

Objectif: Analyser l'évolution de la performance des clubs entre deux saisons (2021-2022 2022-2023) en fonction de leur nombre de recrues durant l'été 2022.

L'objectif est de voir s'il existe une corrélation entre l'activité sur le marché des transferts (quantité de recrutements) et l'évolution du classement d'une saison sur l'autre.

```r
# Regrouper les clubs qui ont recruté et compter
nb_recrues_club <- transfers_clean %>%
  group_by(club_to) %>%
  summarise(nb_recrues = n(), .groups = "drop")

# Préparer les classements par saison
classements_2021_2022 <- team_stats_2021_2022 %>%
  select(Squad, LgRk) %>%
  rename(club = Squad, classement_2022 = LgRk)

classements_2022_2023 <- team_stats_2022_2023 %>%
  select(Squad, LgRk) %>%
  rename(club = Squad, classement_2023 = LgRk)

# Fusionner tout ensemble
evolution_clubs <- classements_2021_2022 %>%
  inner_join(classements_2022_2023, by = "club") %>%
  left_join(nb_recrues_club, by = c("club" = "club_to")) %>%
  mutate(
    nb_recrues = replace_na(nb_recrues, 0),  # certains clubs peuvent avoir 0 recrutement
    evolution_classement = classement_2022 - classement_2023
  )

# Pour la visualisation, créer des classes selon nombre de recrues
evolution_clubs <- evolution_clubs %>%
  mutate(
```

```r
    categorie_recrutement = case_when(
      nb_recrues == 0 ~ "0 recrue",
      nb_recrues <= 3 ~ "1-3 recrues",
      nb_recrues <= 6 ~ "4-6 recrues",
      TRUE ~ "7+ recrues"
    )
  )

# Visualisation : boxplot pour voir la distribution de l'évolution du classement par catégorie de recru
ggplot(evolution_clubs, aes(x = categorie_recrutement, y = evolution_classement, fill = categorie_recru
  geom_boxplot(alpha = 0.7) +
  scale_fill_manual(values = c("lightblue", "skyblue", "blue", "darkblue")) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Impact du recrutement sur l'évolution du classement (2021-2023)",
    subtitle = "Distribution de l'évolution de position en fonction du nombre de recrues estivales",
    x = "Nombre de recrues (catégories)",
    y = "Evolution du classement (positif = amélioration)",
    fill = "Recrutements"
  ) +
  theme_minimal(base_size = 12)
```



Interprétation du graphique: