rapport

Ousmane Tom BECHIR

2025-10-02

1.2 Analyse Descriptive

A. Attributs et Statistiques Descriptives

a. Presentation des differents attributs

a.1. Attributs numeriques

```
str(data_final[vars_numeriques])
```

```
## 'data.frame':    18205 obs. of  13 variables:
##  $ age           : num  28.9 33 21.9 24.9 24.9 ...
##  $ IMC           : num  22.7 23.5 21.7 24.7 23.6 ...
##  $ height_cm     : num  190 175 186 195 176 182 193 188 177 193 ...
##  $ weight_kg     : num  82 72 75 94 73 75 92 86 71 91 ...
##  $ overall_rating: num  91 91 90 90 90 90 90 90 89 89 ...
##  $ potential     : num  91 91 94 92 94 93 90 90 92 89 ...
##  $ value_num     : num  1.16e+08 1.04e+08 1.74e+08 1.57e+08 1.72e+08 ...
##  $ wage_clean    : num  440 350 280 270 340 380 220 170 125 130 ...
##  $ Attaque       : num  84.5 89.2 75.2 83.5 25 ...
##  $ Passe         : num  86 81 83 81 NaN 89 79 82 85 85 ...
##  $ Dribble       : num  64 NaN 50 72 55 NaN 34 57 61 63 ...
##  $ Defense       : num  27 32 32 34 NaN 59 42 33 78 29 ...
##  $ Physique      : num  63 79 64 86 NaN 63 79 68 80 62 ...
```

Parmis les attributs numériques, nous avons des nouvelles attributs qui n'étaient pas dans les données originales que nous avions crée. Un ces attributs présentent quelques valeurs manquantes que nous avions complété en récuperant les valeurs dans les données de `male_players(legacy)` . Malgrès tout, ils restent toujours des données manquantes mais avec une faible fréquences.

b. Statistiques de base

```
cat("\n=== STATISTIQUES SUR L'ÂGE ===\n")
```

```
##
## === STATISTIQUES SUR L'ÂGE ===
```

```
print(summary(age))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.05   22.42   25.83   26.31   29.59   43.80
```

```
# Distribution de l'IMC
cat("\n=== STATISTIQUES SUR L'IMC ===\n")
```

```
##
## === STATISTIQUES SUR L'IMC ===
```

```
print(summary(data_final$IMC))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.02   21.84   22.69   22.70   23.55   29.68
```

```
# Répartition par groupe d'âge
cat("\n=== RÉPARTITION PAR GROUPE D'ÂGE ===\n")
```

```
##
## === RÉPARTITION PAR GROUPE D'ÂGE ===
```

```
print(table(GA))
```

```
## GA
##   <20 20-25 25-30 30-35   >35
##  1428  6452  6154  3341   830
```

```
# Répartition par position catégorisée (club)
cat("\n=== RÉPARTITION PAR POSITION (CLUB) ===\n")
```

```
##
## === RÉPARTITION PAR POSITION (CLUB) ===
```

```
print(table(club_position_cat))
```

```
## club_position_cat
##   DEF   FWD    GK   MID   SUB
##  2627  1078   657  2875 10883
```

```
# Statistiques sur les scores de compétences
cat("\n=== STATISTIQUES SUR LES SCORES DE COMPÉTENCES ===\n")
```

```
##
## === STATISTIQUES SUR LES SCORES DE COMPÉTENCES ===
```

```
cat("Score Attaque:\n")
```

```
## Score Attaque:
```

```
print(summary(data_final$Attaque))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.333  42.750  52.250  50.434  60.000  89.250
```

```
cat("\nScore Physique:\n")
```

```
##
## Score Physique:
```

```
print(summary(data_final$Physique))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   28.00   59.00   66.00   65.35   73.00   91.00    1983
```

```
cat("\nScore Défense:\n")
```

```
##
## Score Défense:
```

```
print(summary(data_final$Defense))
```

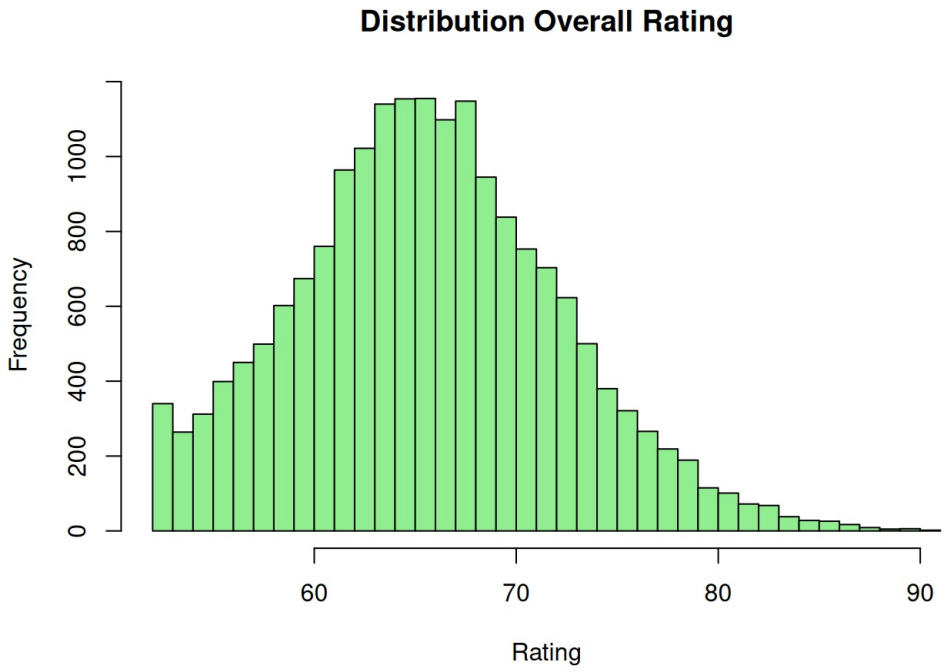
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   34.00   54.00   50.99   64.00   90.00    1983
```

L'analyse des statistiques révèle plusieurs observations importantes sur notre jeu de données :
- **Caractéristiques démographiques** : L'âge moyen des joueurs est de 26,3 ans avec une médiane à 25,8 ans, ce qui indique une population relativement jeune. La majorité des joueurs se situe dans les tranches d'âge 20-25 ans (6 452 joueurs) et 25-30 ans (6 154 joueurs), représentant ensemble plus de 69% de l'effectif total.
L'IMC moyen s'établit à 22,7 kg/m^2, ce qui correspond à un indice de masse corporelle normal et homogène pour des athlètes professionnels.
- **Répartition par position** : On constate une forte proportion de joueurs en position SUB (10 883 joueurs, soit environ 60% du dataset), ce qui suggère que beaucoup de joueurs ne sont pas titulaires. Les milieux de terrain (MID) et défenseurs (DEF) représentent les positions les plus représentées après les remplaçants, avec respectivement 2 875 et 2 627 joueurs.
- **Scores de compétences** : Les scores d'attaque présentent une moyenne de 50,4 sur 100 avec une distribution relativement équilibrée (médiane à 52,3). Les scores physiques (moyenne 65,4) et de défense (moyenne 51,0) montrent des profils plus variés. Cependant, on note la présence de 1 983 valeurs manquantes (environ 11% des données) pour les scores physiques et défensifs, malgré la complétion effectuée à partir du dataset legacy.

c. Visualisation

Distribution de la variable Overall Rating

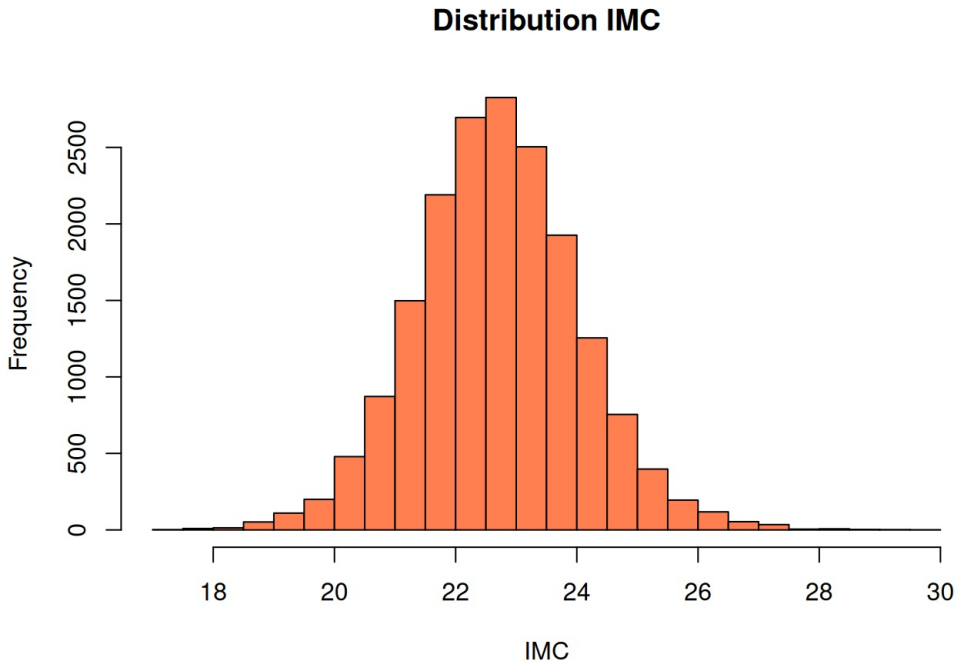
```
hist(data_final$overall_rating, main = "Distribution Overall Rating", xlab = "Rating", col = "lightgreen", breaks = 30)
```



La variable `overall_rating` présente une distribution quasi-normale centrée autour de 65-68. On observe une concentration importante des observations dans l'intervalle [60, 75], ce qui représente environ 80% de l'effectif total. La queue de distribution à droite est relativement courte, indiquant une rareté des valeurs supérieures à 80. Cette forme de distribution est cohérente avec un système de notation bien calibré où la majorité des individus se situent dans une plage moyenne, tandis que les valeurs extrêmes (très faibles ou très élevées) sont moins fréquentes.

Distribution de l'Indice de Masse Corporelle (IMC)

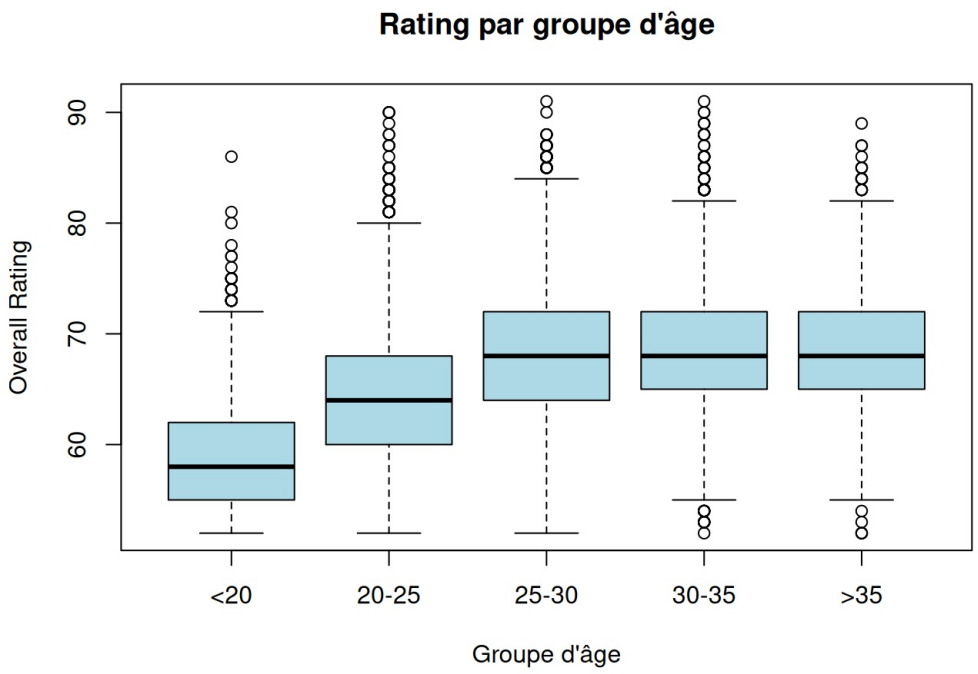
```
hist(data_final$IMC, main = "Distribution IMC", xlab = "IMC", col = "coral", breaks = 30)
```



L'IMC calculé à partir des variables `weight_kg` et `height_cm` montre une distribution symétrique et concentrée autour de 23-24. La quasi-totalité des observations (>95%) se situe dans l'intervalle [21, 25], ce qui correspond à la catégorie "poids normal" selon les standards médicaux. On note très peu de valeurs aberrantes, ce qui suggère une homogénéité importante de cette variable dans notre population d'étude. Cette concentration peut s'expliquer par le fait que les footballeurs professionnels maintiennent une condition physique optimale, avec un IMC dans la fourchette athlétique normale.

Relation entre Overall Rating et Groupe d'Âge

```
boxplot(overall_rating ~ groupe_age, data = data_final, main = "Rating par groupe d'âge",
        xlab = "Groupe d'âge", ylab = "Overall Rating", col = "lightblue")
```



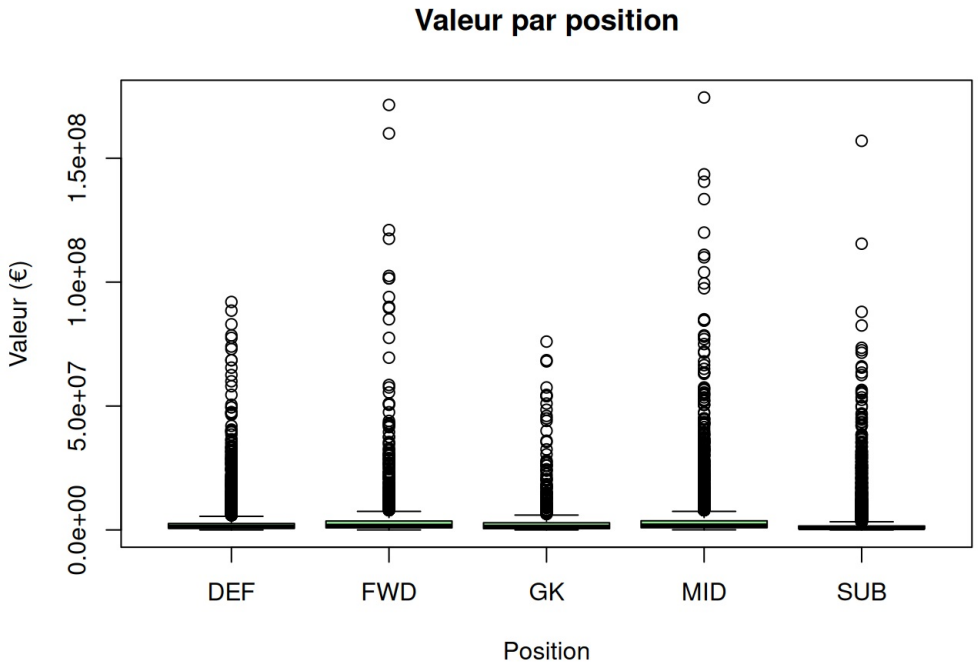
L'analyse par boxplot révèle une relation non-linéaire entre l'âge et le rating :

- **Groupe <20** : Médiane à 58 avec une forte dispersion. Présence de nombreux outliers supérieurs, suggérant un potentiel de développement élevé pour certains individus jeunes.
- **Groupe 20-25** : Progression de la médiane à ~65, avec réduction de la dispersion inter-quartile.
- **Groupe 25-30** : Atteinte du pic de performance (médiane ~68-70), distribution plus resserrée indiquant une stabilisation des compétences.
- **Groupe 30-35** : Maintien du niveau de performance, médiane stable autour de 68.
- **Groupe >35** : Légère diminution mais médiane encore élevée (~67). La dispersion réduite suggère un effet de sélection où seuls les individus performants restent actifs.

Conclusion : On observe une courbe de progression typique avec une phase de croissance (avant 25 ans), un plateau de performance optimale (25-35 ans), puis un léger déclin.

Distribution de la Valeur par Position

```
boxplot(value_num ~ club_position_cat, data = data_final, main = "Valeur par position",
        xlab = "Position", ylab = "Valeur (€)", col = "lightgreen")
```

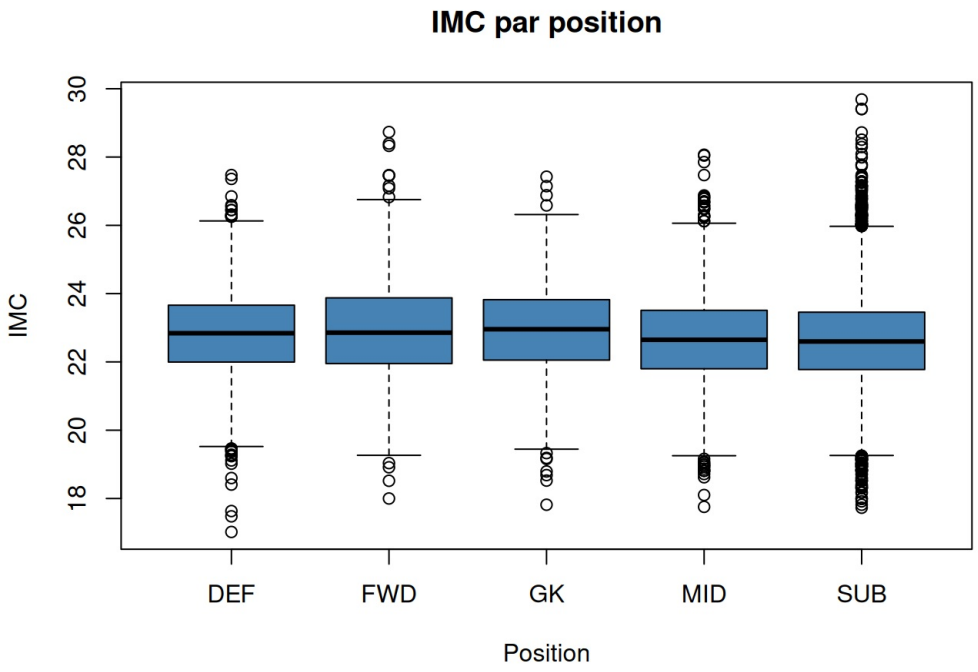


Cette visualisation met en évidence une distribution fortement asymétrique de la variable `value` :

- **Concentration près de zéro** : Pour toutes les catégories de position, la médiane et les quartiles sont très proches de la valeur minimale, indiquant que la majorité des observations ont une faible valeur.
- **Outliers extrêmes** : Présence de valeurs aberrantes très élevées (>100M), particulièrement prononcées pour les positions MID et FWD.
- **Disparité inter-positions** : Les positions offensives (FWD, MID) présentent des valeurs maximales plus élevées que les positions défensives (DEF, GK).
- **Catégorie SUB** : Comme attendu, cette catégorie présente les valeurs les plus faibles avec moins d'outliers.
- **Interprétation statistique** : Cette distribution suggère une forte asymétrie positive (skewness élevé) avec une queue de distribution longue en haut. Cela indique une concentration de la valeur sur un petit nombre d'observations, caractéristique d'un marché où quelques individus d'élite captent une part disproportionnée de la valeur totale.

Relation entre IMC et Position

```
boxplot(IMC ~ club_position_cat, data = data_final, main = "IMC par position",
        xlab = "Position", ylab = "IMC", col = "steelblue")
```



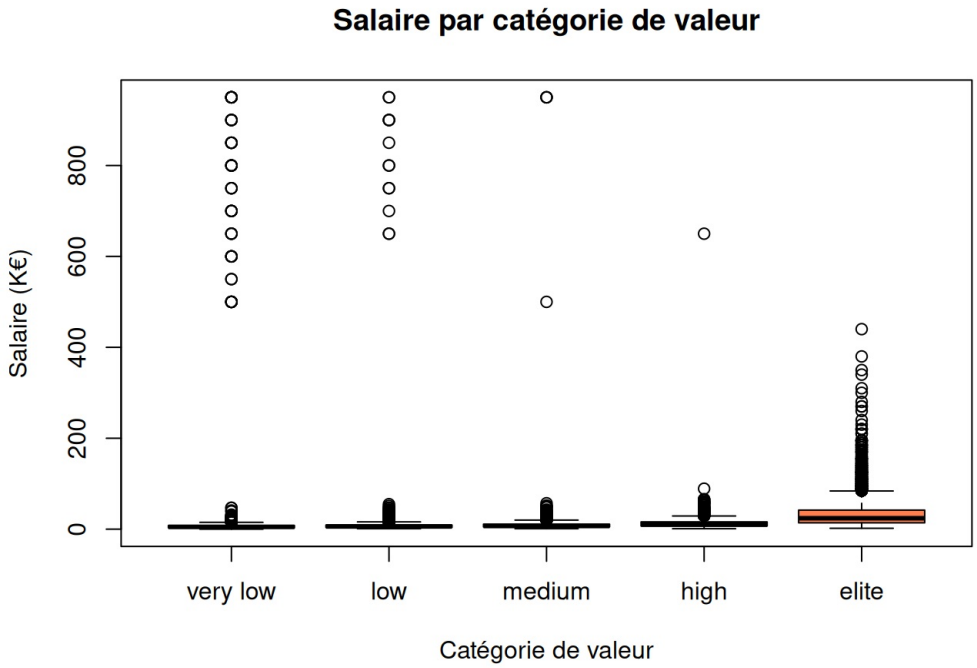
L'analyse comparative de l'IMC selon la position révèle une homogénéité remarquable :

- **Médianes similaires** : Toutes les positions affichent une médiane autour de 22-23, avec des intervalles inter-quartiles très proches.
- **Faible variance inter-groupes** : Les différences observées entre positions sont minimes et probablement non significatives statistiquement.
- **Position GK** : Présente une médiane légèrement supérieure, ce qui peut s'expliquer par des exigences physiques différentes.
- **Outliers symétriques** : Présence d'outliers dans les deux directions (IMC faibles et élevés) pour toutes les positions.

Conclusion : La variable IMC ne semble pas être un facteur discriminant entre les différentes positions. Cette homogénéité suggère que les exigences physiques de base (ratio poids/taille) sont relativement standardisées, quelle que soit la spécialisation.

Relation entre Salaire et Catégorie de Valeur

```
boxplot(wage_clean ~ value_cat, data = data_final, main = "Salaire par catégorie de valeur",
        xlab = "Catégorie de valeur", ylab = "Salaire (K€)", col = "coral")
```



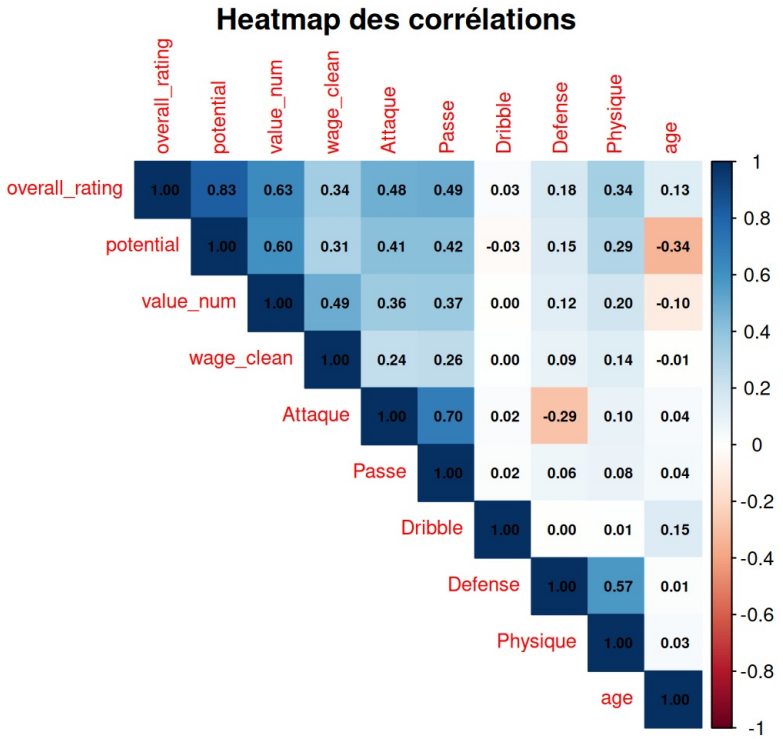
Cette visualisation illustre une relation forte entre les variables `wage` et `value_cat` :

- **Progression monotone** : Le salaire médian augmente de manière monotone avec la catégorie de valeur.
- **Catégories “very low” à “high”** : Salaires médians très faibles (<10K), fortement concentrés près de zéro avec peu de dispersion.
- **Catégorie “elite”** : Changement d’échelle radical avec une médiane à ~50K et une assez forte dispersion. Présence de nombreux outliers dépassant 200K.
- **Asymétrie marquée** : Pour toutes les catégories, on observe une distribution asymétrique positive avec des outliers uniquement vers le haut.

Interprétation : Cette relation confirme une corrélation assez forte entre la valeur marchande et la rémunération. Cependant, la présence d’outliers même dans les catégories basses suggère des inefficiences de marché ou des facteurs non capturés par la variable `value` (ancienneté, réputation, bonus contractuels, etc.). La distribution dans la catégorie “elite” indique également une forte hétérogénéité salariale même parmi les individus les plus valorisés.

B. Analyse de corrélation

```
corrplot(cor_matrix, method = "color", type = "upper",
         addCoef.col = "black", number.cex = 0.6, tl.cex = 0.8,
         title = "Heatmap des corrélations", mar = c(0,0,1,0))
```



La matrice de corrélation révèle plusieurs relations intéressantes entre les variables de notre dataset.

> - **Corrélation entre overall_rating et value_num** : On observe une corrélation positive modérée de 0,63 entre le rating global et la valeur marchande des joueurs. Cette corrélation, bien que significative, n'est pas aussi forte qu'on pourrait l'attendre. Cela suggère que d'autres facteurs influencent la valeur d'un joueur au-delà de sa simple note globale, notamment l'âge (qui présente d'ailleurs une corrélation négative de -0,10 avec la valeur), le potentiel, ou encore la position du joueur.

> - **Intercorrélations entre attributs techniques** : Les scores de compétences techniques montrent des corrélations variables. La relation la plus forte s'observe entre Attaque et Passe (0,70), ce qui est cohérent puisque les joueurs offensifs nécessitent de bonnes capacités de passe. En revanche, certaines corrélations sont faibles voire négatives : Attaque et Défense (-0,29) présentent une corrélation négative notable, reflétant la spécialisation des joueurs selon leur position. Le Dribble montre des corrélations faibles avec les autres attributs (0,05 avec Défense, 0,09 avec Physique), suggérant qu'il s'agit d'une compétence relativement indépendante. Ces résultats confirment que les attributs techniques ne sont pas uniformément corrélés, mais dépendent fortement des profils de poste.

> - **Corrélation entre wage_clean et potential** : La corrélation entre le salaire et le potentiel est de 0,33, ce qui est modéré. Cela indique que le salaire est davantage déterminé par les performances actuelles (corrélation de 0,51 avec overall_rating) que par le potentiel futur. Les clubs semblent donc rémunérer principalement la valeur immédiate plutôt que les perspectives d'évolution.

> - **Corrélation la plus surprenante** : La corrélation qui me surprend le plus est celle entre potential et age (-0,33). Cette corrélation négative modérée signifie que les joueurs plus jeunes ont généralement un potentiel plus élevé, ce qui est logique d'un point de vue footballistique. Cependant, l'intensité de cette corrélation souligne l'importance cruciale de l'âge dans l'évaluation du potentiel d'un joueur. Par ailleurs, la très faible corrélation entre Physique et age (0,03) est également surprenante, car on pourrait s'attendre à ce que les capacités physiques diminuent avec l'âge. Cela suggère que les joueurs professionnels maintiennent un niveau physique élevé tout au long de leur carrière, ou que les données reflètent uniquement les joueurs en activité (biais de survie).

En conclusion, la matrice de corrélation révèle que les performances et la valeur des joueurs sont déterminées par un ensemble complexe de facteurs, où l'âge, la position et les compétences spécifiques jouent des rôles distincts et parfois contradictoires.

C. Prise en main des données

1. L'équipe la plus chère

```
print(equipe_chere %>% select(name, club_position_cat, overall_rating, value_num))
```

```
## # A tibble: 11 × 4
##    name                              club_position_cat overall_rating value_num
##    <chr>                             <chr>                     <dbl>     <dbl>
##  1 Jude Victor William Bellingham -  MID                          90 174500000
##  2 Vini Jr. -                        FWD                          90 171500000
##  3 Kylian Mbappé Lottin -            FWD                          90 160000000
##  4 Florian Richard Wirtz -           MID                          89 143500000
##  5 Lamine Yamal -                    MID                          86 140500000
##  6 Jamal Musiala -                   MID                          88 133500000
##  7 William Alain André Gabriel Salib… DEF                         87  92000000
##  8 Alessandro Bastoni -              DEF                          87  88500000
##  9 Jules Olivier Koundé -            DEF                          86  83000000
## 10 Achraf Hakimi Mouh -              DEF                          86  78500000
## 11 Gianluigi Donnarumma -            GK                           87  76000000
```

2. L'équipe la plus forte

```
print(equipe_forte %>% select(name, club_position_cat, overall_rating, value_num))
```

```
## # A tibble: 11 × 4
##    name                              club_position_cat overall_rating value_num
##    <chr>                             <chr>                     <dbl>     <dbl>
##  1 Mohamed Salah Hamed Ghaly -       MID                          91 104000000
##  2 Jude Victor William Bellingham -  MID                          90 174500000
##  3 Vini Jr. -                        FWD                          90 171500000
##  4 Kylian Mbappé Lottin -            FWD                          90 160000000
##  5 Virgil van Dijk -                 DEF                          90  77500000
##  6 Florian Richard Wirtz -           MID                          89 143500000
##  7 Alisson -                         GK                           89  54500000
##  8 Kevin De Bruyne -                 MID                          89  63500000
##  9 Antonio Rüdiger -                 DEF                          88  62500000
## 10 William Alain André Gabriel Salib… DEF                          87  92000000
## 11 Alessandro Bastoni -              DEF                          87  88500000
```

3. Comparaison

```
cat("Joueurs en commun:", length(joueurs_communs), "sur 11\n")
```

```
## Joueurs en commun: 6 sur 11
```

```
if(length(joueurs_communs) > 0) {
  cat("Noms:", paste(joueurs_communs, collapse = ", "), "\n")
} else {
  cat("Aucun joueur en commun\n")
}
```

```
## Noms: Jude Victor William Bellingham -, Vini Jr. -, Kylian Mbappé Lottin -, Florian Richard Wirtz -, William A
lain André Gabriel Saliba -, Alessandro Bastoni -
```

Nous observons que les deux équipes ne sont pas formées totalement par les même joueurs. Avec seulement 6 joueurs en commun sur 11, cela signifie que 5 joueurs diffèrent entre l'équipe la plus chère et l'équipe la plus forte (soit environ 45% de l'effectif). **Conclusion par rapport à l'analyse de corrélation** : Ce résultat confirme parfaitement notre observation dans l'analyse de la matrice de corrélation, où nous avions constaté une corrélation modérée de 0,63 entre `overall_rating` et `value_num` . Cette corrélation, bien que significative, n'était pas parfaite, et nous avions conclu que d'autres facteurs influençaient la valeur marchande au-delà de la seule performance actuelle. Ces résultat confirme bien notre précedente conclusion.

4. Une équipe par ligue ou par pays

```
cat("\n=== ÉQUIPE PREMIER LEAGUE ===\n")
```

```
##
## === ÉQUIPE PREMIER LEAGUE ===
```

```
print(equipe_premier_league %>% select(name, club_position_cat, overall_rating, club_name))
```

```
## # A tibble: 11 × 4
##    name                              club_position_cat overall_rating club_name
##    <chr>                             <chr>                     <dbl> <chr>
##  1 Mohamed Salah Hamed Ghaly -       MID                          91 Liverpool
##  2 Virgil van Dijk -                 DEF                          90 Liverpool
##  3 Alisson -                         GK                           89 Liverpool
##  4 Kevin De Bruyne -                 MID                          89 Manchest…
##  5 Bukayo Saka -                     FWD                          88 Arsenal
##  6 Cole Jermaine Palmer -            MID                          87 Chelsea
##  7 William Alain André Gabriel Salib… DEF                         87 Arsenal
##  8 Declan Rice -                     MID                          87 Arsenal
##  9 Rúben Dias -                      DEF                          86 Manchest…
## 10 Alexander Isak -                  FWD                          86 Newcastl…
## 11 Trent John Alexander-Arnold -     DEF                          86 Liverpool
```

```
cat("\n=== ÉQUIPE FRANCE ===\n")
```

```
##
## === ÉQUIPE FRANCE ===
```

```
print(equipe_france %>% select(name, club_position_cat, overall_rating, club_name))
```

```
## # A tibble: 11 × 4
##    name                              club_position_cat overall_rating club_name
##    <chr>                             <chr>                     <dbl> <chr>
##  1 Kylian Mbappé Lottin -            FWD                          90 Real Mad…
##  2 William Alain André Gabriel Salib… DEF                         87 Arsenal
##  3 Masour Ousmane Dembélé -          FWD                          87 Paris Sa…
##  4 Mike Peterson Maignan -           GK                           87 AC Milan
##  5 Jules Olivier Koundé -            DEF                          86 FC Barce…
##  6 Theo Bernard François Hernández - DEF                          86 AC Milan
##  7 Michael Akpovie Olise -           MID                          85 FC Bayer…
##  8 Ibrahima Konaté -                 DEF                          85 Liverpool
##  9 N'Golo Kanté -                    MID                          85 Al Ittih…
## 10 Aurélien Djani Tchouameni -       MID                          84 Real Mad…
## 11 Adrien Rabiot-Provost -           MID                          83 Olympiqu…
```

5. Créer une équipe prometteuse (jeunes à fort potentiel)

```
print(equipe_prometteuse %>%
       select(name, club_position_cat, age, overall_rating, potential, marge_progression))
```

```
## # A tibble: 11 × 6
##    name       club_position_cat   age overall_rating potential marge_progression
##    <chr>      <chr>             <dbl>          <dbl>     <dbl>              <dbl>
##  1 Andre Gar… DEF                17.5             59        84                 25
##  2 Filip Öhm… DEF                17.4             57        80                 23
##  3 Dylan Leo… DEF                17.8             53        76                 23
##  4 Oliver Jo… FWD                19.0             53        76                 23
##  5 Semm Rend… DEF                17.5             62        84                 22
##  6 Matias Si… MID                18.2             61        82                 21
##  7 Maxloren … MID                17.5             64        84                 20
##  8 Harry Joh… MID                18.2             63        83                 20
##  9 Mason Mel… FWD                17.7             62        82                 20
## 10 Matthew Y… GK                 18.5             59        79                 20
## 11 Travis En… MID                19.6             58        78                 20
```

```
cat("Âge moyen:", round(mean(equipe_prometteuse$age, na.rm = TRUE), 2), "ans\n")
```

```
## Âge moyen: 18.07 ans
```

```
cat("Marge de progression moyenne:", round(mean(equipe_prometteuse$marge_progression, na.rm = TRUE), 2), "\n")
```

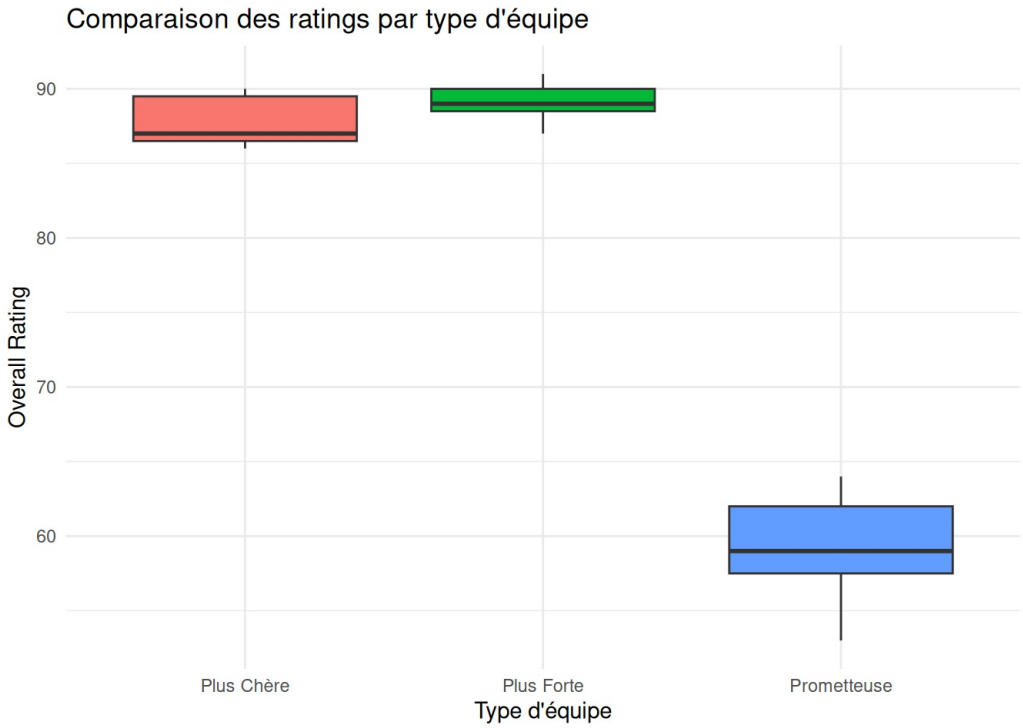
```
## Marge de progression moyenne: 21.55
```

Nous avons créer une équipe prometteuse en se basant sur la marge de progression et sur l'e potentiel'âge (`marge_progression = potential - overall_rating`), nous avons prioriser les joueurs avec une grande marge de progression (plus haut potentiel).

Quelques visualisations comparatives des équipes

Comparaison des rating par type d'équipe

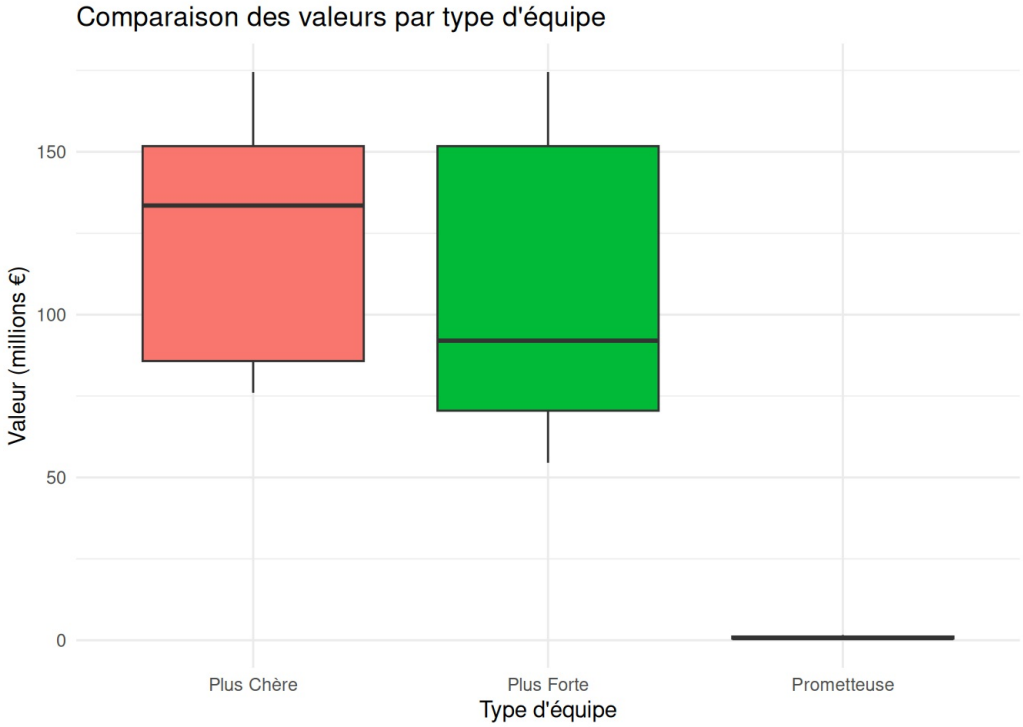
```
ggplot(equipes_comparison, aes(x = Type, y = overall_rating, fill = Type)) +
  geom_boxplot() +
  labs(title = "Comparaison des ratings par type d'équipe",
       x = "Type d'équipe", y = "Overall Rating") +
  theme_minimal() +
  theme(legend.position = "none")
```



Comme observer précedement, nous avons des valeurs de médiane de `Overall Rating` plus élevés pour l'équipe la plus chère et l'équipe la plus forte, avec une médiane un peu plus au dessus pour l'équipe la plus forte ce qui est logique et vient même de l'essence de l'équipe la plus forte. Cette légère variation vient encore confirmer notre conclusion sur la correlation entre la valeur marchand et le overall rating. Et comme attendu nous avons une assez faible dispertion de la valeur du overall rating pour l'équipe la plus forte. Quant à l'équipe la plus prometteuse, nous avons une médiane rélativement basse par rapport aux autres équipes, ce qui est tout à fait logique et s'explique par le jeune âge des joueurs de cette équipe.

Comparaison des valeurs par type d'équipe

```
ggplot(equipes_comparison %>% filter(!is.na(value_num)),
       aes(x = Type, y = value_num/1e6, fill = Type)) +
  geom_boxplot() +
  labs(title = "Comparaison des valeurs par type d'équipe",
       x = "Type d'équipe", y = "Valeur (millions €)") +
  theme_minimal() +
  theme(legend.position = "none")
```



C'est tout à fais le résultat attendu du point de vue de la construction des équipes. Nous avons des valeurs marchandes très élevés pour l'équipe la plus chère par rapport aux autres équipes, avec une médiane à environ 130 millions d'euros, tandis que l'équipe la plus forte présente une médiane légèrement inférieure à 90 millions d'euros. Cette différence significative confirme que les joueurs les plus chers ne sont pas nécessairement les plus performants actuellement.

L’équipe prometteuse, composée de jeunes talents, affiche des valeurs marchandes proches de zéro, ce qui est logique puisque ces joueurs n’ont pas encore atteint leur pic de performance malgré leur fort potentiel de progression. La distribution des valeurs pour les équipes “Plus Chère” et “Plus Forte” montre également une certaine dispersion, avec des interquartiles étendus, suggérant une hétérogénéité dans les profils des joueurs sélectionnés. Ces observations renforcent une fois de plus l’analyse de corrélation précédente : la `valeur marchande` intègre plusieurs dimensions (performance actuelle, potentiel, âge, durée de contrat) alors que le rating reflète principalement les compétences instantanées du joueur.