Analyse exploratoire

29/05/2019

```
transfers <- read.csv("../../data/clean_transfers.csv")
```

Le dataset contient les 250 plus gros transferts des saisons 2000-2001 à 2018-2019 (à moins de 10 transferts près par saison). Après nettoyage des données, nous disposons de sept prédicteurs, quatre qualitatives et trois quantitatives discrètes :

```
sapply(transfers, class)
```

```
##     Position          Age  League_from    League_to       Season
##     "factor"    "integer"     "factor"     "factor"     "factor"
## Market_value Transfer_fee
##    "integer"    "integer"
```

Donnons-nous une première idée de la répartition des différentes variables :

```
summary(transfers)
```

```
##        Position         Age                League_from
##  Defender  :1122   Min.   :15.00   1.Bundesliga  : 265
##  Forward   :1923   1st Qu.:22.00   LaLiga        : 418
##  Goalkeeper: 180   Median :24.00   Ligue 1       : 428
##  Midfielder:1475   Mean   :24.34   Premier League: 608
##                    3rd Qu.:27.00   Serie A       : 602
##                    Max.   :35.00   NA's          :2379
##                    NA's   :1
##           League_to          Season      Market_value
##  1.Bundesliga  : 422   2001-2002: 250   Min.   :    50000
##  LaLiga        : 525   2008-2009: 250   1st Qu.:  3500000
##  Ligue 1       : 397   2013-2014: 250   Median :  6000000
##  Premier League:1256   2016-2017: 250   Mean   :  8622469
##  Serie A       : 739   2006-2007: 249   3rd Qu.: 10000000
##  NA's          :1361   2009-2010: 249   Max.   :120000000
##                        (Other)  :3202   NA's   :1260
##   Transfer_fee
##  Min.   :   825000
##  1st Qu.:  4000000
##  Median :  6500000
##  Mean   :  9447586
##  3rd Qu.: 10820000
##  Max.   :222000000
##
```

Première analyse :

- Position : **les attaquants sont les plus représentés**, suivis des milieux de terrain et des défenseurs. Il y a dix fois plus d'attaquants que de gardiens
- Age : une médiane à 24 ans. 50 % entre 22 et 27 ans
- League_from : **Premier League (Angleterre) et Serie A (Italie)** à égalité, puis LaLiga (Espagne) et Ligue 1 (France) au même niveau également, avant la Bundesliga. La moitié des valeurs mises à `NA`

pendant le nettoyage, autrement dit : la moitié des transferts ont été faits depuis une des cinq ligues majeures

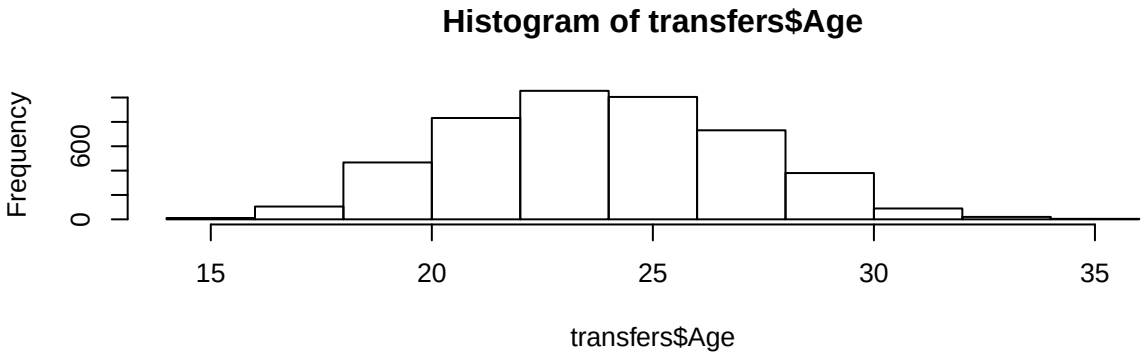
- League_to : Premier League loin devant, puis Serie A, puis les trois autres en dessous. Moins de valeurs `NA` : 65 % des transferts avaient une League_to appartenant aux cinq ligues majeures. **Plus grande variété de League_from que de League_to**. Hypothèse : beaucoup de clubs peuvent vendre des joueurs cher, mais peu de clubs peuvent acheter des joueurs chers, et ces derniers sont dans les cinq ligues majeures.
- Season : RAS
- Market_value : médiane à 6 M, 75 % entre 3,5 M et 10 M, max à 120 M.
- Transfer_fee : médiane à 6,5 M, 75 % entre 4 M et 10,8 M, max à 222 M. **Un décalage net avec Market_value** que l'on peut confirmer avec le boxplot suivant :

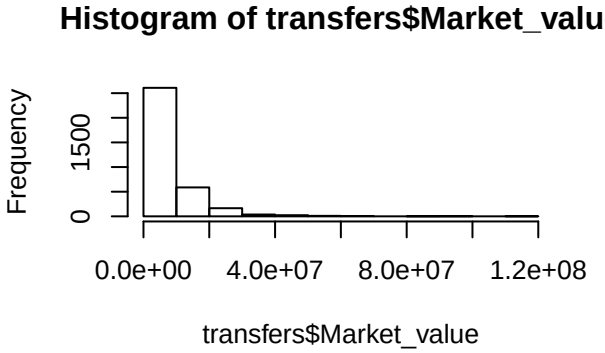
Variables quantitatives

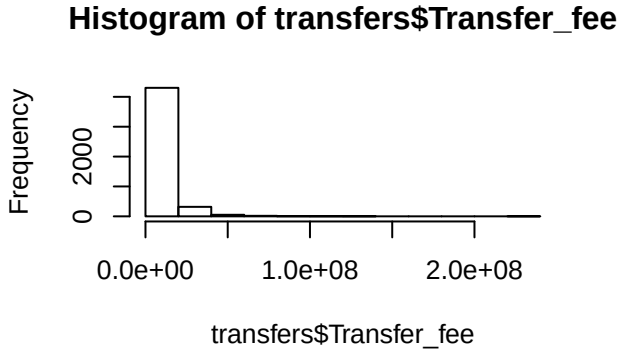
Analyse univariée

L'âge suit une loi normale mais pas Market_value et Transfer_fee

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(transfers$Age)
hist(transfers$Market_value)
hist(transfers$Transfer_fee)
```



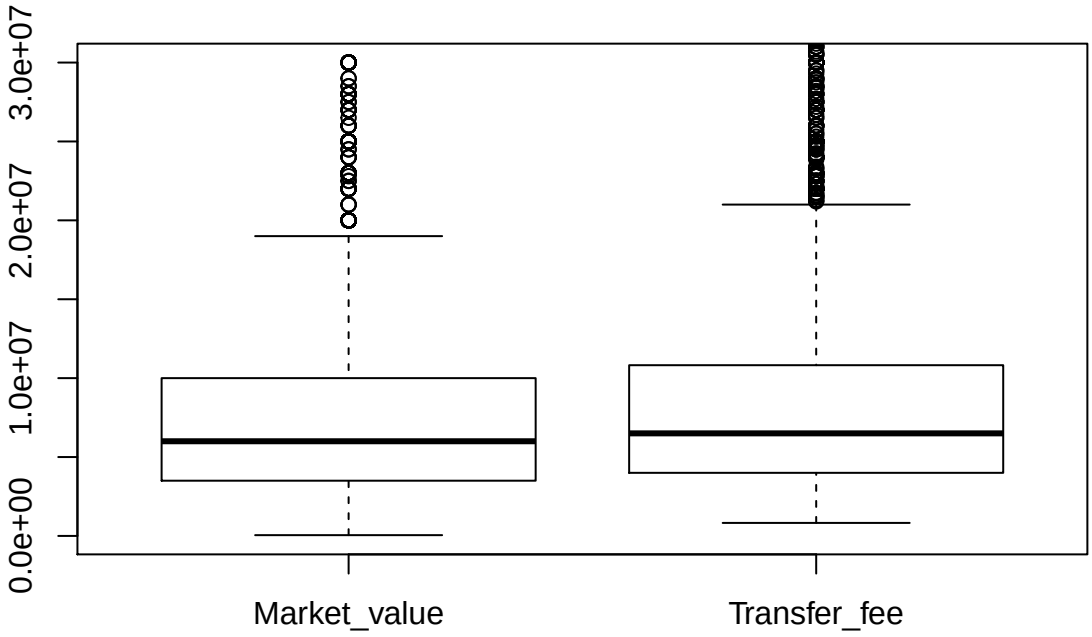




Transfer_fee plus élevé que Market_value et beaucoup de valeurs aberrantes

Transfer_fee semble toujours plus élevé que la Market_value, et il y a **plus de variance au-dessus de la médiane** dans les deux cas. On note aussi beaucoup de valeurs aberrantes, on doit donc davantage se baser sur la médiane que la moyenne.

```
boxplot(transfers$Market_value, transfers$Transfer_fee, ylim = c(30000, 30000000),
        names = c("Market_value", "Transfer_fee"))
```

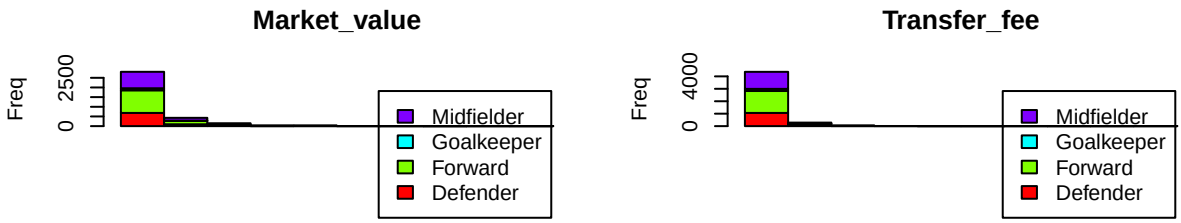


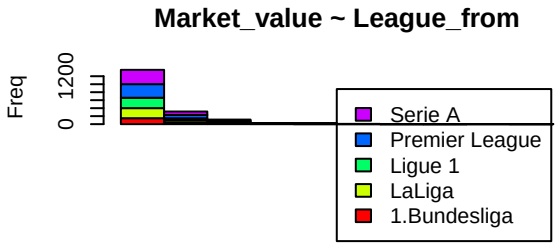
Analyse bidimensionnelle

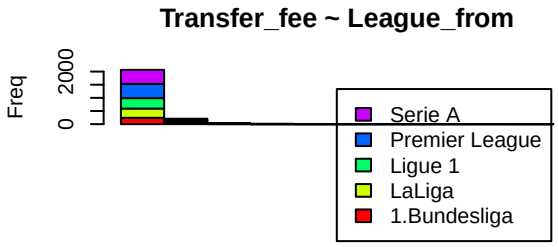
Les attaquants sont vendus plus cher, les clubs anglais et espagnols dépensent le plus

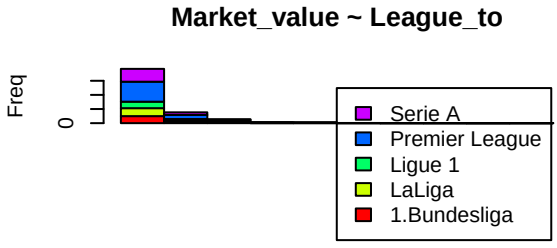
```
hist.factor <- function (var_quanti, var_quali, title, return) {
  inter <- seq(min(var_quanti, na.rm = T), max(var_quanti, na.rm = T),
               by = (max(var_quanti, na.rm = T) - min(var_quanti, na.rm = T))/10)
  hists <- c()
  for (mod in levels(var_quali)) {
    h = hist(plot = F, var_quanti[var_quali == mod], breaks = inter)
    hists <- rbind(hists, h$counts)
  }

  if (return == "barplot") {
      barplot(hists, space = 0, legend = levels(var_quali), ylab = "Freq",
              main = title, col=rainbow(nlevels(var_quali)))
  } else {
    hists
  }
}

# Transfer_fee selon position
old.par <- par(mfrow=c(3, 2))
```

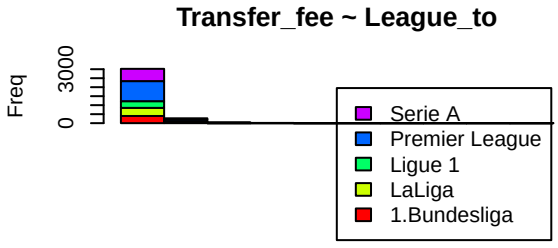
```r
hist.factor(transfers$Market_value, transfers$Position, "Market_value", "barplot")
hist.factor(transfers$Transfer_fee, transfers$Position, "Transfer_fee", "barplot")
hist.factor(transfers$Market_value, transfers$League_from, "Market_value ~ League_from", "barplot")
hist.factor(transfers$Transfer_fee, transfers$League_from, "Transfer_fee ~ League_from", "barplot")
hist.factor(transfers$Market_value, transfers$League_to, "Market_value ~ League_to", "barplot")
hist.factor(transfers$Transfer_fee, transfers$League_to, "Transfer_fee ~ League_to", "barplot")
```







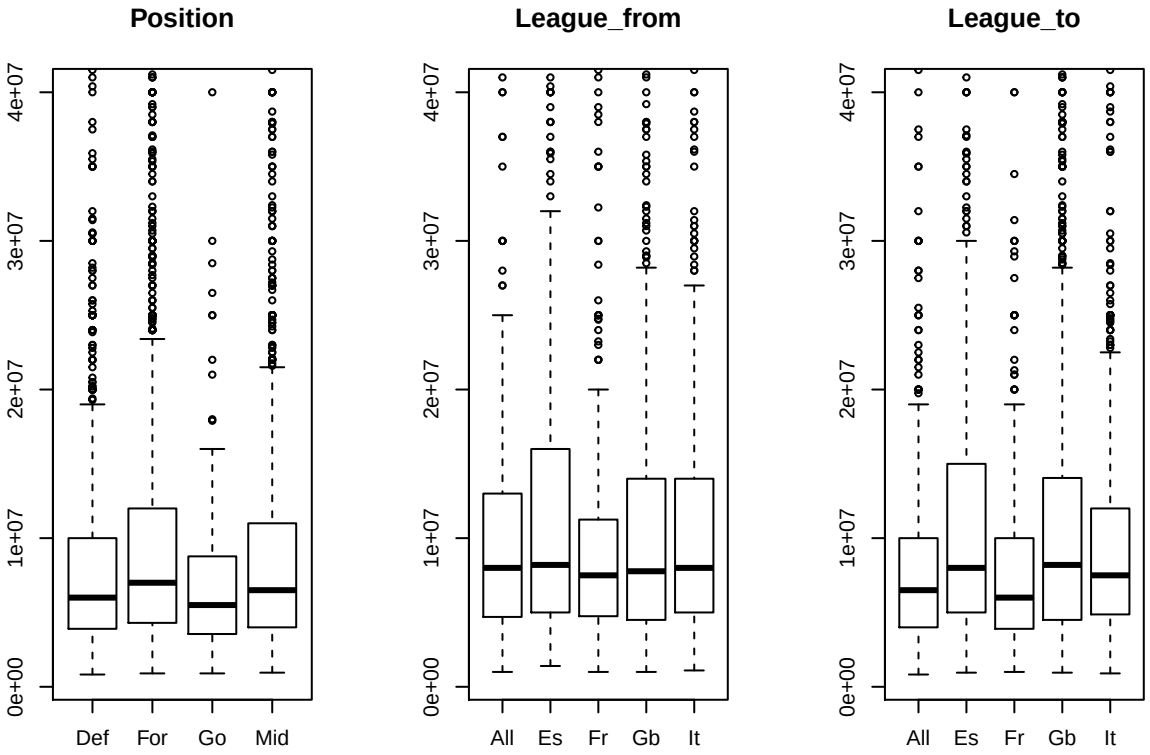





```r
par(old.par)
```

On constate une **symétrie entre les graphiques deux à deux**, et entre League_from et League_to**. On retrouve la répartition donnée par la fonction `summary`. On peut confirmer avec des boxplots :

```r
old.par <- par(mfrow=c(1, 3))
boxplot(transfers$Transfer_fee~transfers$Position, ylim = c(700000,40000000),
        main = "Position", names = c("Def", "For", "Go", "Mid"))
boxplot(transfers$Transfer_fee~transfers$League_from, ylim = c(700000,40000000),
        main = "League_from", names = c("All", "Es", "Fr", "Gb", "It"))
boxplot(transfers$Transfer_fee~transfers$League_to, ylim = c(700000,40000000),
        main = "League_to", names = c("All", "Es", "Fr", "Gb", "It"))
```

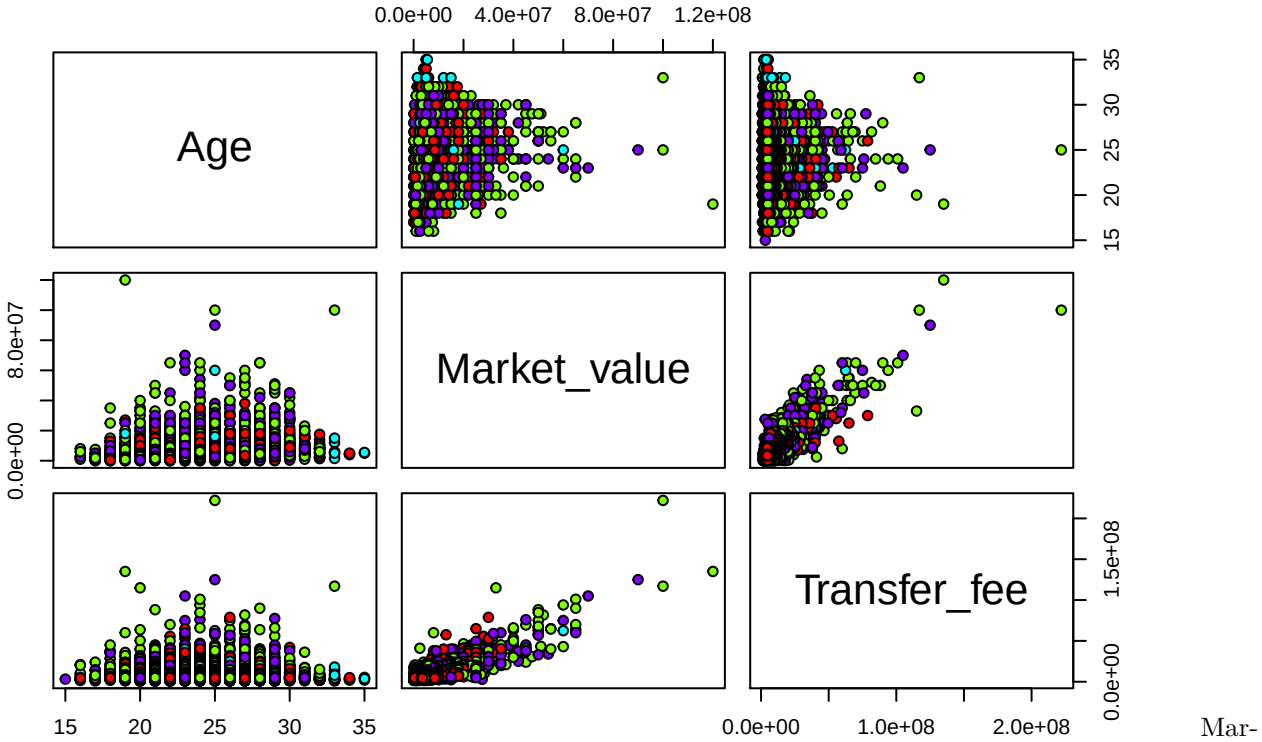



```
par(old.par)
```

Homogénéité entre les ligues sur les médianes des prix de vente, mais moins sur les prix d'achat. Les clubs espagnols et anglais achètent plus cher. Par ailleurs, les attaquants sont bien vendus plus cher.

Graphique matriciel sur Age, Market_value et Transfer_fee

```
pairs(transfers[c(2,6,7)], pch = 21, bg = rainbow(4)[transfers$Position])
```



Market_value dépend linéairement de Transfer_fee mais on le comprend : si un joueur a été acheté cher, il y a des chances pour que son prix sur le marché était déjà élevé. La linéarité est moins évidente dans l'autre sens, laissant supposer une certaine irrationnalité avec **des prix d'achat beaucoup plus élevés que la valeur sur le marché**.

Confirmation avec étude de corrélation :

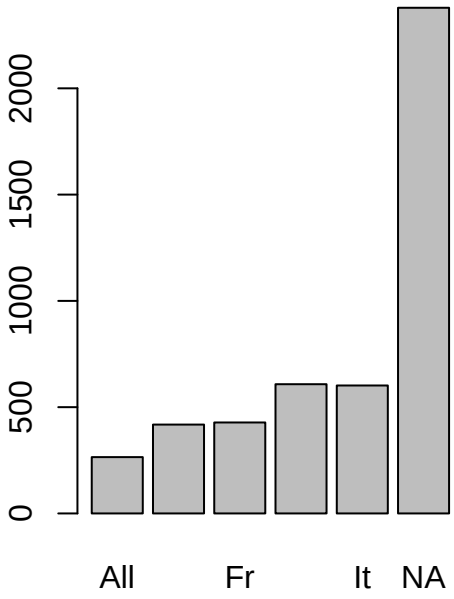
```
cor(transfers$Market_value, transfers$Transfer_fee, use = "complete.obs")
```

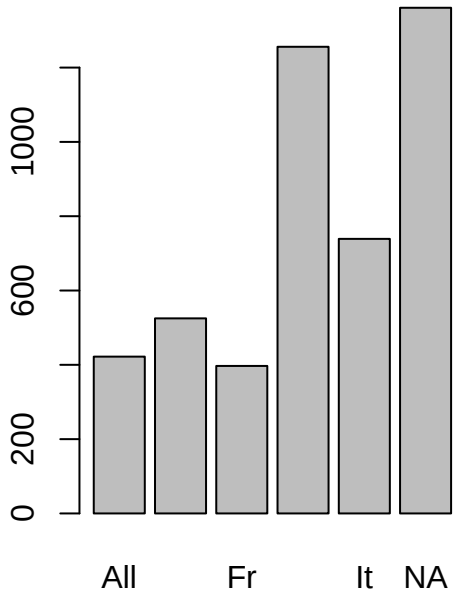
```
## [1] 0.8305728
```

Corrélation positive et proche de 1.

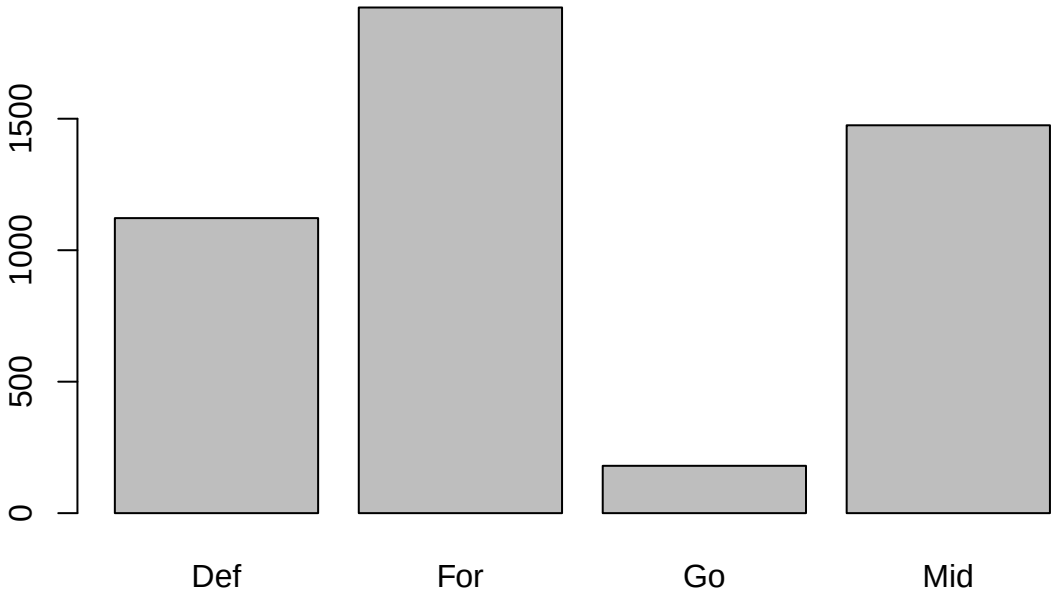
Variables qualitatives

```
old.par <- par(mfrow=c(1, 2))
names.arg <- c("All", "Es", "Fr", "Gb", "It", "NA")
barplot(summary(transfers$League_from), names.arg = names.arg)
barplot(summary(transfers$League_to), names.arg = names.arg)
```






```r
par(old.par)
barplot(summary(transfers$Position), names.arg = c("Def", "For", "Go", "Mid"))
```



Synthèse :

- position : les attaquants sont ceux qui sont vendus et achetés les plus chers
- âge : suit une loi normale
- League_from : pas uniformément répartie, valeurs aberrantes élevées. Certaines ligues mettent sur le marché des prix très élevés, notamment l'Espagne
- League_to : idem que League_from, mais valeurs plus élevées. L'Espagne et l'Allemagne sont les ligues qui achètent aux prix les plus élevés
- League_from et League_to : corrélées positivement