

# Analyse exploratoire

## Taille du dataset

### Nombre de transferts enregistrés

```
nrow(transfers)
```

```
## [1] 4700
```

Le dataset répertorie donc 4700 transferts. Il est sensé contenir les 250 transferts les plus élevés pour chaque saison des dix-neuf dernières saisons :

```
length(unique(transfers$Season))
```

```
## [1] 19
```

Or :

```
19 * 250
```

```
## [1] 4750
```

Nous ne disposons donc pas exactement de 250 transferts par saison :

```
freq <- table(transfers$Season)
sort(freq[freq != 250])
```

```
##
## 2003-2004 2002-2003 2017-2018 2010-2011 2018-2019 2014-2015 2005-2006
##      242      244      244      245      245      246      247
## 2000-2001 2004-2005 2007-2008 2012-2013 2015-2016 2006-2007 2009-2010
##      248      248      248      248      248      249      249
## 2011-2012
##      249
```

Il manque des transferts pour 15 saisons sur 19 mais ce n'est jamais plus de 8 transferts sur 250, ce qui ne devrait pas fausser les mesures de l'influence de la saison que l'on pourra faire par la suite.

### Nombre de prédicteurs

```
ncol(transfers)
```

```
## [1] 10
```

```
colnames(transfers)
```

```
## [1] "Name"          "Position"      "Age"           "Team_from"
## [5] "League_from"   "Team_to"       "League_to"     "Season"
## [9] "Market_value" "Transfer_fee"
```

## Nettoyage des données

### Vérification des classes des prédicteurs

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

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

On convertit la variable “Name” au type `character`, utilisé en R pour représenter les `string`. On pourrait dire que le nom est un facteur mais la variable prend tellement de valeurs que cela semble plus pertinent d’en faire un type `character`. On ne le fait pas pour les équipes et les ligues, qui prennent moins de valeurs différentes et peuvent a priori avoir une influence sur le prix des joueurs, contrairement à leur nom.

```
transfers$Name <- as.character(transfers$Name)
class(transfers$Name)
```

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

On convertit la variable “Season” en facteur ordonné :

```
# we use the already alphabetical order of seasons
transfers$Season <- as.ordered(transfers$Season)
```

### Vérification cohérence des données

#### Valeurs manquantes ou nulles

```
summary(transfers)
```

```
##      Name      Position      Age
## Length:4700   Centre-Forward :1218   Min.   :15.00
## Class :character Centre-Back   : 714   1st Qu.:22.00
## Mode  :character Central Midfield : 487   Median :24.00
##      Attacking Midfield: 426   Mean   :24.34
##      Defensive Midfield: 411   3rd Qu.:27.00
##      Right Winger      : 305   Max.   :35.00
##      (Other)           :1139   NA's   :1
```

```
##      Team_from      League_from      Team_to
## Inter      : 68      Premier League: 608      Inter      : 97
## Spurs       : 63      Serie A         : 602      Chelsea     : 96
## Juventus    : 59      Ligue 1         : 428      Man City    : 94
## Chelsea     : 57      LaLiga          : 418      Spurs       : 93
## FC Porto    : 56      1.Bundesliga  : 265      Juventus    : 87
## Liverpool: 56      Série A         : 199      Liverpool: 85
## (Other)     :4341      (Other)         :2180      (Other)     :4148
##      League_to      Season      Market_value
## Premier League:1256  2001-2002: 250      Min.       : 50000
## Serie A          : 739  2008-2009: 250      1st Qu.: 3500000
## LaLiga           : 525  2013-2014: 250      Median    : 6000000
## 1.Bundesliga     : 422  2016-2017: 250      Mean      : 8622469
## Ligue 1          : 397  2006-2007: 249      3rd Qu.: 10000000
## Premier Liga     : 328  2009-2010: 249      Max.      :120000000
## (Other)          :1033  (Other)    :3202      NA's      :1260
##      Transfer_fee
## Min.       : 825000
## 1st Qu.: 4000000
## Median    : 6500000
## Mean      : 9447586
## 3rd Qu.: 10820000
## Max.      :222000000
##
```

Seule la colonne “Market\_value” contient des valeurs nulles, à raison de 1260 sur 4700 soit 27 %.

```
# extract rows where Market_value is na
null_market_value <- transfers[is.na(transfers$Market_value) == T,]

# make a contingency table by season
cont <- table(null_market_value$Season)
cont
```

```
##
## 2000-2001 2001-2002 2002-2003 2003-2004 2004-2005 2005-2006 2006-2007
##      248      250      244      242      189      28      20
## 2007-2008 2008-2009 2009-2010 2010-2011 2011-2012 2012-2013 2013-2014
##      13       7       2       4       1       2       2
## 2014-2015 2015-2016 2016-2017 2017-2018 2018-2019
##       1       0       1       3       3
```

```
# proportion of the first five seasons
sum(cont[1:5])
```

```
## [1] 1173
```

```
sum(cont[1:5]) / sum(cont)
```

```
## [1] 0.9309524
```

Les cinq premières saisons concentrent l’essentiel des valeurs manquantes.

## Valeurs incohérentes

On s'intéresse aux ligues :

```
levels(transfers$League_from)
```

```
## [1] " Argentina"
## [3] " Brazil"
## [5] " Canada"
## [7] " China"
## [9] " Croatia"
## [11] " Denmark"
## [13] " England"
## [15] " France"
## [17] " Iran"
## [19] " Latvia"
## [21] " Moldova"
## [23] " Peru"
## [25] " Qatar"
## [27] " Russia"
## [29] " Scotland"
## [31] " Slovakia"
## [33] " Spain"
## [35] " Tunisia"
## [37] " United Arab Emirates"
## [39] " Uruguay"
## [41] "1.Bundesliga"
## [43] "1.HNL"
## [45] "1.Liga gr. 1"
## [47] "2ª B - Grupo I"
## [49] "3.Liga"
## [51] "Allsvenskan"
## [53] "Botola Pro"
## [55] "Challenge League"
## [57] "Championship"
## [59] "Eliteserien"
## [61] "First Division"
## [63] "J1 - 2nd Stage"
## [65] "J2 League"
## [67] "K League 1"
## [69] "LaLiga"
## [71] "League One"
## [73] "Liga 1 - Championship group"
## [75] "Liga MX Apertura"
## [77] "Liga NOS"
## [79] "Ligue 1"
## [81] "Ligue I Pro"
## [83] "NB I."
## [85] "Premier League"
## [87] "Premiership"
## [89] "Primera B Nacional"
## [91] "Primera División"
## [93] "Proximus League"
## [95] "Rel. Ligue 1"

" Australia"
" Bulgaria"
" Chile"
" Colombia"
" Czech Republic"
" Ecuador"
" Finland"
" Ghana"
" Korea, South"
" Mexico"
" Paraguay"
" Portugal"
" Romania"
" Saudi Arabia"
" Serbia"
" South Africa"
" Sweden"
" Ukraine"
" United States"
" Venezuela"
"1.Division"
"1.Lig"
"2.Bundesliga"
"2ª B - Grupo III"
"A Grupa - Championship gr."
"Auf-/Abstiegsrunde NLA/NLB"
"Bundesliga"
"Championnat National"
"Ekstraklasa"
"Eredivisie"
"HET Liga"
"J1 League"
"Jupiler Pro League"
"Korean FA Cup"
"LaLiga2"
"Liga 1"
"Liga Águila II"
"Liga MX Clausura"
"Ligat ha'Al"
"Ligue 2"
"MLS"
"OBOS-ligaen"
"Premier Liga"
"Primavera B"
"Primera Div. Apertura"
"Professional League"
"Regionalliga Nord"
"Second Division (bis 03/04)"
```

```
## [97] "Segunda División"      "Serie A"
## [99] "Série A"               "Serie A Segunda Etapa"
## [101] "Serie B"               "Série B"
## [103] "Serie C - A"           "Serie C - B"
## [105] "Stars League"          "Super League"
## [107] "Süper Lig"             "Superettan"
## [109] "SuperLiga"            "Superligaen"
## [111] "Superligaen Championship round" "Torneo Final"
## [113] "Torneo Inicial"        "U18 Premier League"
## [115] "U19 Eredivisie"        "UAE Gulf League"
## [117] "Virsliga"              "Vysheyskaya Liga"
```

Les 40 premières valeurs sont des valeurs de pays. Les autres valeurs sont bien des ligues. On regarde combien de transferts ont une valeur de pays et non de ligue dans une des deux colonnes de ligues :

```
pays_League_from <- levels(transfers$League_from)[1:40]
pays_League_to <- levels(transfers$League_to)[1:24]
pays <- unique(c(pays_League_from, pays_League_to))
nrow(subset(transfers, League_from %in% pays | League_to %in% pays))
```

```
## [1] 445
```

444 transferts sur 4699 sont concernés, soit presque 10 %. Nous décidons néanmoins de nous en débarrasser :

```
transfers <- subset(transfers, !(League_from %in% pays | League_to %in% pays))
nrow(transfers)
```

```
## [1] 4255
```

## Synthèse

Nous avons changé le type de la conomme “Name”, remarqué que 27 % des valeurs de la colonne “Market\_value” étaient manquantes, principalement sur les cinq premières saisons, supprimé l’unique transfert où l’âge du joueur était nul, et supprimé les 10 % de transferts donc la colonne “League\_from” ou “League\_to” avait pour valeur un pays.

## Première étude des données

```
length(levels(transfers$Team_from))
```

```
## [1] 570
```

```
length(levels(transfers$Team_to))
```

```
## [1] 325
```

Les 4699 transferts se répartissent en 570 clubs vendeurs et 325 clubs acheteurs. Cette différence pourrai s’expliquer par le fait que certains clubs vendent cher des joueurs dont ils ont fait augmenter la valeur après les avoir acheté à faible prix. Il y a une concentration des mêmes clubs acheteurs dans les transferts les plus élevés.