

# Nettoyage du dataset

## 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"
```

## 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 "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
## Alex      : 8  Centre-Forward :1218  Min.   :15.00
## Fernando  : 7  Centre-Back    : 714  1st Qu.:22.00
## Peter Crouch : 7  Central Midfield : 487  Median :24.00
## Adriano    : 6  Attacking Midfield: 426  Mean   :24.34
## Alberto Gilardino : 6  Defensive Midfield: 411  3rd Qu.:27.00
## Carlos T\303\251vez: 6  Right Winger      : 305  Max.   :35.00
## (Other)    :4660  (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  S\303\251rie 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 une grande quantité de NA, 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.

## Simplification du dataset

### Ligues

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"

" 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"
```

## [41]	"1.Bundesliga"	"1.Division"
## [43]	"1.HNL"	"1.Lig"
## [45]	"1.Liga gr. 1"	"2.Bundesliga"
## [47]	"2\302\252 B - Grupo I"	"2\302\252 B - Grupo III"
## [49]	"3.Liga"	"A Grupa - Championship gr."
## [51]	"Allsvenskan"	"Auf-/Abstiegsrunde NLA/NLB"
## [53]	"Botola Pro"	"Bundesliga"
## [55]	"Challenge League"	"Championnat National"
## [57]	"Championship"	"Ekstraklasa"
## [59]	"Eliteserien"	"Eredivisie"
## [61]	"First Division"	"HET Liga"
## [63]	"J1 - 2nd Stage"	"J1 League"
## [65]	"J2 League"	"Jupiler Pro League"
## [67]	"K League 1"	"Korean FA Cup"
## [69]	"LaLiga"	"LaLiga2"
## [71]	"League One"	"Liga 1"
## [73]	"Liga 1 - Championship group"	"Liga MX Apertura"
## [75]	"Liga MX Clausura"	"Liga NOS"
## [77]	"Liga \303\201guila II"	"Ligat ha'Al"
## [79]	"Ligue 1"	"Ligue 2"
## [81]	"Ligue I Pro"	"MLS"
## [83]	"NB I."	"OBOS-ligaen"
## [85]	"Premier League"	"Premier Liga"
## [87]	"Premiership"	"Primavera B"
## [89]	"Primera B Nacional"	"Primera Div. Apertura"
## [91]	"Primera Divisi\303\263n"	"Professional League"
## [93]	"Proximus League"	"Regionalliga Nord"
## [95]	"Rel. Ligue 1"	"Second Division (bis 03/04)"
## [97]	"Segunda Divisi\303\263n"	"Serie A"
## [99]	"Serie A Segunda Etapa"	"Serie B"
## [101]	"Serie C - A"	"Serie C - B"
## [103]	"Stars League"	"Super League"
## [105]	"SuperLiga"	"Superettan"
## [107]	"Superligaen"	"Superligaen Championship round"
## [109]	"S\303\251rie A"	"S\303\251rie B"
## [111]	"S\303\274per Lig"	"Torneo Final"
## [113]	"Torneo Inicial"	"U18 Premier League"
## [115]	"U19 Eredivisie"	"UAE Gulf League"
## [117]	"Virsliga"	"Vysheyskaya Liga"

Exemple des vérifications faites sur les noms de ligue ambigus : "Série A" correspond à la première division brésilienne et "Serie A" à l'italienne, "Bundesliga" à l'autrichienne et "1. Bundesliga" à l'allemande. On remarque aussi des valeurs de pays. Dans la plupart des cas, ces valeurs correspondent à des divisions inférieures des pays correspondants.

Hypothèse : la plupart des transferts impliquent un des cinq grands championnats : anglais, espagnol, allemand, français et italien :

```
# we define a vector containing the five main leagues names
leagues.main <- ( c("1.Bundesliga", "LaLiga", "Ligue 1", "Premier League", "Serie A"))

# we extract transfers between clubs of the five main leagues
transfers.main <- transfers[transfers$League_from %in% leagues.main & transfers$League_to %in% leagues.main]

# proportion
```

```
nrow(transfers.main)
```

```
## [1] 1965
```

```
nrow(transfers.main) / nrow(transfers)
```

```
## [1] 0.4180851
```

42 % des transferts impliquent uniquement des clubs des cinq grands championnats, ce qui représente tout de même près de 2000 transferts. Nous décidons de mettre toutes les autres valeurs de ligue à NA. Ainsi, nous diminuons le nombre de modalités des ligues :

```
# we set to NA all the leagues that are not in the five main leagues
transfers$League_from[! transfers$League_from %in% leagues.main] <- NA
transfers$League_to[! transfers$League_to %in% leagues.main] <- NA
```

```
# we reset the levels of the leagues columns
transfers$League_from <- factor(transfers$League_from)
transfers$League_to <- factor(transfers$League_to)
```

## Equipes et noms des joueurs

On retire les colonnes noms et les équipes car elles proposent plusieurs valeurs beaucoup qu'à notre avis n'influencent pas beaucoup notre analyse :

```
transfers <- subset(transfers, select = -c(Name, Team_from, Team_to))
```

## Positions

On regroupe en quatre catégories : gardiens, défenseurs, milieux, attaquants :

```
levels(transfers$Position)
```

```
## [1] "Attacking Midfield" "Central Midfield" "Centre-Back"
## [4] "Centre-Forward" "Defender" "Defensive Midfield"
## [7] "Forward" "Goalkeeper" "Left Midfield"
## [10] "Left Winger" "Left-Back" "Midfielder"
## [13] "Right Midfield" "Right Winger" "Right-Back"
## [16] "Second Striker" "Sweeper"
```

```
levels(transfers$Position) <- gsub(".*Back$|Sweeper", "Defender", levels(transfers$Position))
levels(transfers$Position) <- gsub(".*Forward$|Second Striker|.*Winger$", "Forward",
                                   levels(transfers$Position))
levels(transfers$Position) <- gsub(".*Midfield$", "Midfielder", levels(transfers$Position))
levels(transfers$Position)
```

```
## [1] "Midfielder" "Defender" "Forward" "Goalkeeper"
```

## Synthèse

- la variable "Saison" a été convertie en facteur ordonné
- il manque la valeur "Market\_value" pour 27 % des transferts et les cinq premières saisons concentrent plus de 90 % de ces valeurs manquantes

- pour les ligues, 41 % des transferts du dataset, soit près de 2000, ont été effectués entre deux clubs des cinq plus grands championnats. Nous avons mis toutes les autres valeurs de ligue à **NA**, pour ne plus avoir que cinq modalités de ligues.
- nous avons supprimé les colonnes “Name”, “Team\_from” et “Team\_to”
- les positions ont été regroupées en quatre catégories : gardiens, défenseurs, milieux, attaquants