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"
```

Vérification cohérence des données

Valeurs manquantes

```
summary(transfers)
```

```
##      Name                          Position         Age
##  Length:4700        Centre-Forward    :1218   Min.   : 0.00
##  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
##      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 %.

```
na <- transfers[is.na(transfers$Market_value) == T,]
table(na$Season)
```

```
##
## 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
```

L’essentiel des valeurs manquantes se concenttre donc dans les cinq premières saisons. Nous n’allons pas supprimer les transferts correspondants car cela nous priverait d’une proportion non négligeable des données pour les calculs non basés sur la “Market_value”. Pour les calculs l’intégrant, on devra veiller à exclure les transferts des cinq premières saisons.