



# **Analyse des performances sportives offensives en ligue 1 Uber Eats**

*Nettoyage de la base de donnée*

KONE Hassan-Ahmed Sékou

## Abstract

L'objectif de ce document est de fournir un exemple de nettoyage d'une base de données avec le logiciel de programmation R afin de la rendre exploitable pour l'analyse. Pour ce faire, nous baserons notre démarche sur les performances sportives individuelles des joueurs de 'Ligue 1 Uber Eats' lors de la saison 2021/2022 à la 30ème journée. Ce guide a été entièrement rédigé en langage R markdown pour une production en format pdf. Les données et la définition des attributs sont disponibles en cliquant sur <**fbref**>. Egalement, le script '.R' et le fichier '.Rmd' sont disponibles en cliquant sur <**Gitbub**>.

**Keywords:** *classes des variables, données manquantes, variable continue, variable discrète, discrétisation, filtre*

# Contents

<b>1</b>	<b>Organisation du travail</b>	<b>2</b>
1.1	Création du répertoire de travail . . . . .	2
1.2	Liste des packages chargés . . . . .	2
<b>2</b>	<b>Manipulation des bases de données</b>	<b>3</b>
2.1	Chargement des bases de données à fuisonner . . . . .	3
2.2	Fusion des bases de données . . . . .	4
2.3	Renomination des variables . . . . .	4
2.4	Classe des variables . . . . .	7
2.4.1	Modification de la variable age dans le bon format . . . . .	8
2.4.2	Modification de la classe des autres variables . . . . .	8
2.5	Relocalisation des variables . . . . .	10
<b>3</b>	<b>Analyse exploratoire de la bdd</b>	<b>11</b>
3.1	Structure des données . . . . .	11
3.2	Données manquantes . . . . .	12
3.3	distribution des variables . . . . .	15
3.3.1	Variables qualitatives . . . . .	15
3.3.2	Variables continues . . . . .	16
3.3.2.1	Normalité . . . . .	16
3.3.2.2	Evolution du nombre de buts en fonction des autres variables continues . . . . .	24

# Chapter 1

## Organisation du travail

### 1.1 Création du répertoire de travail

```
setwd("~/Desktop/analyse performance joueur ligue 1")  
getwd()
```

```
## [1] "/Users/hassan-ahmedsekoukone/Desktop/analyse performance joueur ligue 1"
```

### 1.2 Liste des packages chargés

```
library(dplyr)  
library(tidyr)  
library(lubridate)  
library(tsoutliers)  
library(readxl)  
library(tidyverse)  
library(stringr)  
library(patchwork)  
library(lattice)  
library(DataExplorer)  
library(VIM)  
library(summarytools)  
library("funModeling")  
library(GGally)  
library(ISLR)  
library(leaps)  
library(knitr)  
library(kableExtra)
```

## Chapter 2

# Manipulation des bases de données

### 2.1 Chargement des bases de données à fuisonner

Dans un document ‘.Rmd’, toujours inclure le chemin du répertoire dans le chunk pour importer ou exporter des éléments de RStudio. Pour un script ‘.R’, pas besoin.

```
setwd("~/Desktop/analyse performance joueur ligue 1")

tirs <- read_excel("perfomance_joueurs.xlsx",
                  sheet = "tirs")

prepa_tirs_buts <- read_excel("perfomance_joueurs.xlsx",
                             sheet = "preparation tirs et buts")

temps_jeu<- read_excel("perfomance_joueurs.xlsx",
                      sheet = "temps de jeu")
```

```
kable(head(tirs[,1:8]),format = "latex")
```

Joueur	Nation	Pos	Équipe	Âge	Naissance	90	Buts
Yunis Abdelhamid	ma MAR	DF	Reims	34-192	1987	26.1	1
Salis Abdul Samed	gh GHA	MT	Clermont Foot	22-013	2000	23.6	1
Laurent Abergel	fr FRA	MT	Lorient	29-066	1993	25.2	0
Charles Abi	fr FRA	AT	Saint-Étienne	21-361	2000	0.5	0
Matthis Abline	fr FRA	AT	Rennes	19-011	2003	1.1	0
Martin Adeline	fr FRA	MTAT	Reims	18-127	2003	3.9	0

-dimension de la base de données : (586, 19)

```
kable(head(prepa_tirs_buts[,1:8]),format = "latex")
```

Joueur	Équipe	AMT	AMT90	PassJeu...5	PassArr...6	Drib...7	Tirs...8
Yunis Abdelhamid	Reims	15	0.57	10	0	2	1
Salis Abdul Samed	Clermont Foot	34	1.44	24	0	0	4
Laurent Abergel	Lorient	42	1.67	29	2	0	0
Charles Abi	Saint-Étienne	0	0.00	0	0	0	0
Matthis Abline	Rennes	1	0.87	1	0	0	0
Martin Adeline	Reims	8	2.05	7	0	1	0

-dimension de la base de données : (586, 18)

```
kable(head(temps_jeu[, 1:8]), format = "latex")
```

Joueur	Équipe	MJ	Min	Mn/MJ	Min%	Titulaire	Mn/Débuté
Yunis Abdelhamid	Reims	27	2353	87	87.1	27	87
Salis Abdul Samed	Clermont Foot	26	2128	82	78.8	25	84
Laurent Abergel	Lorient	26	2269	87	84.0	26	87
Charles Abi	Saint-Étienne	1	45	45	1.7	1	45
Matthis Abline	Rennes	7	103	15	3.8	1	45
Mohamed Achi	Nantes	0	NA	NA	NA	0	NA

-dimension de la base de données : (716, 18)

## 2.2 Fusion des bases de données

```
df= merge(temps_jeu,prepa_tirs_buts,by=c("Joueur","Équipe"),all.x = T)
df= merge(df,tirs,by=c("Joueur","Équipe"),all.x = T)
kable(head(df[, 1:8]), format = "latex")
```

Joueur	Équipe	MJ	Min	Mn/MJ	Min%	Titulaire	Mn/Débuté
Abdel Jalil Medioub	Bordeaux	9	537	60	19.9	5	88
Abdou Diallo	Paris S-G	12	877	73	32.5	10	86
Abdoulaye Bakayoko	Saint-Étienne	2	180	90	6.7	2	90
Abdoulaye Bamba	Angers	2	4	2	0.1	0	NA
Abdoulaye Sylla	Nantes	1	67	67	2.5	1	67
Abdu Conté	Troyes	10	823	82	30.5	10	82

-dimension de la base de données : (716, 51)

## 2.3 Renomination des variables

- Nom des variables

```
names(df)
```

```
## [1] "Joueur"
## [3] "MJ"
## [5] "Mn/MJ"
## [7] "Titulaire"
## [9] "Compl"
## [11] "Mn/Remp"
## [13] "PPM"
## [15] "BE"
## [17] "+/-90"
## [19] "AMT"
## [21] "PassJeu...5"
## [23] "Drib...7"
## [25] "Ftp...9"

"Équipe"
"Min"
"Min%"
"Mn/Débuté"
"Remp"
"RempNE"
"BT"
"+/-"
"Sur/En dehors du terrain"
"AMT90"
"PassArr...6"
"Tirs...8"
"Déf...10"
```

## [27] "AMB"	"AMB90"
## [29] "PassJeu...13"	"PassArr...14"
## [31] "Drib...15"	"Tirs...16"
## [33] "Ftp...17"	"Déf...18"
## [35] "Nation"	"Pos"
## [37] "Âge"	"Naissance"
## [39] "90"	"Buts"
## [41] "Tirs"	"TC"
## [43] "TC%"	"Tir/90"
## [45] "TC/90"	"B/Tir"
## [47] "B/TC"	"Dist"
## [49] "CF"	"PénM"
## [51] "PénT"	

Nous devons renommer la majorité des variables pour une meilleure lecture.

```
df= rename(df, age="Âge",
           equipe="Équipe",
           minutes_jouees_90="90",
           Mn_MJ="Mn/MJ",
           pourcentage_TC="TC%",
           Tir_90="Tir/90",
           TC_90="TC/90",
           B_Tir="B/Tir",
           Mn_Debute="Mn/Débuté",
           buts_marques_net_avec_joueur="+/-",
           buts_marques_net_avec_joueur_par_match="+/-90",
           Sur_En_dehors_du_terrain="Sur/En dehors du terrain",
           PassJeu_tirs="PassJeu...5",
           PassArr_tirs="PassArr...6",
           Drib_tirs="Drib...7",
           tirs_tirs="Tirs...8",
           Ftp_tirs="Ftp...9",
           Mn_Remp="Mn/Remp",
           pourcentage_min="Min%",
           def_tirs="Déf...10",
           passjeu_buts="PassJeu...13",
           passarr_buts="PassArr...14",
           Drib_buts="Drib...15",
           tirs_buts="Tirs...16",
           Ftp_buts="Ftp...17",
           Def_buts="Déf...18",
           B_TC="B/TC",
           PenM="PénM",
           PenT="PénT")
```

- Pour obtenir le nom des variables en minuscule.

```
colnames(df)= str_to_lower(colnames(df))
colnames(df)
```

```
## [1] "joueur"
```

```

## [2] "equipe"
## [3] "mj"
## [4] "min"
## [5] "mn_mj"
## [6] "pourcentage_min"
## [7] "titulaire"
## [8] "mn_debute"
## [9] "compl"
## [10] "remp"
## [11] "mn_remp"
## [12] "rempne"
## [13] "ppm"
## [14] "bt"
## [15] "be"
## [16] "buts_marques_net_avec_joueur"
## [17] "buts_marques_net_avec_joueur_par_match"
## [18] "sur_en_dehors_du_terrain"
## [19] "amt"
## [20] "amt90"
## [21] "passjeu_tirs"
## [22] "passarr_tirs"
## [23] "drib_tirs"
## [24] "tirs_tirs"
## [25] "ftp_tirs"
## [26] "def_tirs"
## [27] "amb"
## [28] "amb90"
## [29] "passjeu_buts"
## [30] "passarr_buts"
## [31] "drib_buts"
## [32] "tirs_buts"
## [33] "ftp_buts"
## [34] "def_buts"
## [35] "nation"
## [36] "pos"
## [37] "age"
## [38] "naissance"
## [39] "minutes_jouees_90"
## [40] "buts"
## [41] "tirs"
## [42] "tc"
## [43] "pourcentage_tc"
## [44] "tir_90"
## [45] "tc_90"
## [46] "b_tir"
## [47] "b_tc"
## [48] "dist"
## [49] "cf"
## [50] "penm"
## [51] "pent"

```



## 2.4 Classe des variables

- Variables dans un format inadéquat. En particulier la variable 'age'.

```
str(df)
```

```
## 'data.frame':    716 obs. of  51 variables:
## $ joueur          : chr  "Abdel Jalil Medioub" "Abdou Diall
## $ equipe          : chr  "Bordeaux" "Paris S-G" "Saint-Étie
## $ mj              : num  9 12 2 2 1 10 15 1 24 0 ...
## $ min             : num  537 877 180 4 67 ...
## $ mn_mj           : num  60 73 90 2 67 82 22 6 78 NA ...
## $ pourcentage_min : chr  "19.9" "32.5" "6.7" "0.1" ...
## $ titulaire       : num  5 10 2 0 1 10 2 0 21 0 ...
## $ mn_debute       : num  88 86 90 NA 67 82 79 NA 87 NA ...
## $ compl           : num  4 8 2 0 0 6 1 0 17 0 ...
## $ remp            : num  4 2 0 2 0 0 13 1 3 0 ...
## $ mn_remp         : num  24 11 NA 2 NA NA 13 6 17 NA ...
## $ rempne          : num  16 10 1 13 14 0 13 1 0 18 ...
## $ ppm             : chr  "1.33" "2.50" "0.00" "1.50" ...
## $ bt              : num  11 23 1 0 1 7 8 0 48 NA ...
## $ be              : num  15 9 3 0 2 12 5 0 19 NA ...
## $ buts_marques_net_avec_joueur : num  -4 14 -2 0 -1 -5 3 0 29 NA ...
## $ buts_marques_net_avec_joueur_par_match: chr  "-0.67" "+1.44" "-1.00" "0.00" ...
## $ sur_en_dehors_du_terrain : chr  "+0.49" "+0.35" "-0.21" "+0.33" ...
## $ amt             : num  0 5 1 1 0 9 8 0 38 NA ...
## $ amt90           : chr  "0.00" "0.51" "0.50" "22.50" ...
## $ passjeu_tirs    : num  0 5 0 1 0 7 7 0 30 NA ...
## $ passarr_tirs    : num  0 0 0 0 0 0 0 0 1 NA ...
## $ drib_tirs       : num  0 0 0 0 0 0 1 0 6 NA ...
## $ tirs_tirs       : num  0 0 0 0 0 0 0 0 1 NA ...
## $ ftp_tirs        : num  0 0 1 0 0 1 0 0 0 NA ...
## $ def_tirs        : num  0 0 0 0 0 1 0 0 0 NA ...
## $ amb             : num  0 1 0 0 0 2 1 0 7 NA ...
## $ amb90           : chr  "0.00" "0.10" "0.00" "0.00" ...
## $ passjeu_buts    : num  0 1 0 0 0 2 1 0 6 NA ...
## $ passarr_buts    : num  0 0 0 0 0 0 0 0 0 NA ...
## $ drib_buts       : num  0 0 0 0 0 0 0 0 1 NA ...
## $ tirs_buts       : num  0 0 0 0 0 0 0 0 0 NA ...
## $ ftp_buts        : num  0 0 0 0 0 0 0 0 0 NA ...
## $ def_buts        : num  0 0 0 0 0 0 0 0 0 NA ...
## $ nation          : chr  "dz ALG" "sn SEN" "fr FRA" "ci CIV
## $ pos             : chr  "DF" "DF" "DF" "ATDF" ...
## $ age             : chr  "24-223" "25-339" "19-114" "31-348
## $ naissance       : num  1997 1996 2002 1990 2000 ...
## $ minutes_jouees_90 : chr  "6.0" "9.7" "2.0" "0.0" ...
## $ buts            : num  0 0 0 0 0 0 2 0 3 NA ...
## $ tirs            : num  3 2 0 0 0 3 7 0 17 NA ...
## $ tc              : num  0 1 0 0 0 1 3 0 7 NA ...
## $ pourcentage_tc  : chr  "0.0" "50.0" NA NA ...
## $ tir_90          : chr  "0.50" "0.21" "0.00" "0.00" ...
## $ tc_90           : chr  "0.00" "0.10" "0.00" "0.00" ...
## $ b_tir           : chr  "0.00" "0.00" NA NA ...
```

```
## $ b_tc : chr NA "0.00" NA NA ...
## $ dist : chr "22.1" "14.0" NA NA ...
## $ cf : num 0 0 0 0 0 0 0 0 0 NA ...
## $ penm : num 0 0 0 0 0 0 0 0 0 NA ...
## $ pent : num 0 0 0 0 0 0 0 0 0 NA ...
```

## 2.4.1 Modification de la variable age dans le bon format

```
age=rename(data.frame(joueur = df$joueur
                      ,equipe=df$equipe,
                      str_split_fixed(df$age, "-", 2))
           ,age_annee=X1,age_jours=X2)
age = update_columns(age, c("age_annee","age_jours"),
                    as.numeric)
age = mutate(age,age=age_annee+(age_jours/365))
age=select(age, joueur, equipe, age)
kable(head(age,3),format = 'latex')
```

joueur	equipe	age
Abdel Jalil Medioub	Bordeaux	24.61096
Abdou Diallo	Paris S-G	25.92877
Abdoulaye Bakayoko	Saint-Étienne	19.31233

- 'df' sans l'ancienne variable 'age' située à la place 37.

```
df=df[, -37]
```

- 'df' avec la nouvelle variable 'age' créée dans le bon format.

```
df=merge(df, age, by=c('joueur', 'equipe'), all.x = T)
str(df$age)
```

```
## num [1:716] 24.6 25.9 19.3 32 22 ...
```

## 2.4.2 Modification de la classe des autres variables

```
df = update_columns(df,
                    c("mj", "min", "mn_mj",
                      "pourcentage_min", "titulaire",
                      "tirs_tirs", "mn_debute", "compl",
                      "remp", "mn_remp", "rempne", "ppm",
                      "bt", "be", "buts_marques_net_avec_joueur",
                      "buts_marques_net_avec_joueur_par_match",
                      "sur_en_dehors_du_terrain", "amt90",
                      "passjeu_tirs", "passarr_tirs",
                      "drib_tirs", "tirs", "ftp_tirs", "def_tirs",
                      "amb", "amb90", "passjeu_buts", "passarr_buts",
                      "drib_buts", "tirs_buts", "ftp_buts",
```

```

        "def_buts", "minutes_jouees_90",
        "buts", "tc", "pourcentage_tc",
        "tir_90", "tc_90", "b_tir", "b_tc",
        "dist", "cf", "penm", "pent", "buts")
    , as.numeric)

df = update_columns(df, c("joueur", "equipe", "nation", "pos"
    , "naissance")
    , as.factor)

str(df)

```

```

## 'data.frame':    716 obs. of  51 variables:
## $ joueur          : Factor w/ 699 levels "Abdel Jalil Medico
## $ equipe          : Factor w/ 20 levels "Angers","Bordeaux"
## $ mj              : num  9 12 2 2 1 10 15 1 24 0 ...
## $ min             : num  537 877 180 4 67 ...
## $ mn_mj           : num  60 73 90 2 67 82 22 6 78 NA ...
## $ pourcentage_min : num  19.9 32.5 6.7 0.1 2.5 30.5 12.1 0.
## $ titulaire       : num  5 10 2 0 1 10 2 0 21 0 ...
## $ mn_debute       : num  88 86 90 NA 67 82 79 NA 87 NA ...
## $ compl           : num  4 8 2 0 0 6 1 0 17 0 ...
## $ remp            : num  4 2 0 2 0 0 13 1 3 0 ...
## $ mn_remp         : num  24 11 NA 2 NA NA 13 6 17 NA ...
## $ rempne          : num  16 10 1 13 14 0 13 1 0 18 ...
## $ ppm             : num  1.33 2.5 0 1.5 0 1.2 1.27 0 2.29 N
## $ bt              : num  11 23 1 0 1 7 8 0 48 NA ...
## $ be              : num  15 9 3 0 2 12 5 0 19 NA ...
## $ buts_marques_net_avec_joueur : num  -4 14 -2 0 -1 -5 3 0 29 NA ...
## $ buts_marques_net_avec_joueur_par_match: num  -0.67 1.44 -1 0 -1.34 -0.55 0.83 0
## $ sur_en_dehors_du_terrain : num  0.49 0.35 -0.21 0.33 -1.58 -0.12 0
## $ amt             : num  0 5 1 1 0 9 8 0 38 NA ...
## $ amt90           : num  0 0.51 0.5 22.5 0 0.98 2.2 0 1.82
## $ passjeu_tirs    : num  0 5 0 1 0 7 7 0 30 NA ...
## $ passarr_tirs    : num  0 0 0 0 0 0 0 0 1 NA ...
## $ drib_tirs       : num  0 0 0 0 0 0 1 0 6 NA ...
## $ tirs_tirs       : num  0 0 0 0 0 0 0 0 1 NA ...
## $ ftp_tirs        : num  0 0 1 0 0 1 0 0 0 NA ...
## $ def_tirs        : num  0 0 0 0 0 1 0 0 0 NA ...
## $ amb             : num  0 1 0 0 0 2 1 0 7 NA ...
## $ amb90           : num  0 0.1 0 0 0 0.22 0.28 0 0.34 NA ...
## $ passjeu_buts    : num  0 1 0 0 0 2 1 0 6 NA ...
## $ passarr_buts    : num  0 0 0 0 0 0 0 0 0 NA ...
## $ drib_buts       : num  0 0 0 0 0 0 0 0 1 NA ...
## $ tirs_buts       : num  0 0 0 0 0 0 0 0 0 NA ...
## $ ftp_buts        : num  0 0 0 0 0 0 0 0 0 NA ...
## $ def_buts        : num  0 0 0 0 0 0 0 0 0 NA ...
## $ nation          : Factor w/ 67 levels "ar ARG","at AUT",.
## $ pos             : Factor w/ 10 levels "AT","ATDF","ATMT",
## $ naissance       : Factor w/ 23 levels "1983","1984",...: 1
## $ minutes_jouees_90 : num  6 9.7 2 0 0.7 9.1 3.6 0.1 20.9 NA
## $ buts            : num  0 0 0 0 0 0 2 0 3 NA ...
## $ tirs            : num  3 2 0 0 0 3 7 0 17 NA ...
## $ tc              : num  0 1 0 0 0 1 3 0 7 NA ...

```

```
## $ pourcentage_tc      : num  0 50 NA NA NA 33.3 42.9 NA 41.2 NA
## $ tir_90              : num  0.5 0.21 0 0 0 0.33 1.93 0 0.81 NA
## $ tc_90               : num  0 0.1 0 0 0 0.11 0.83 0 0.34 NA ..
## $ b_tir              : num  0 0 NA NA NA 0 0.29 NA 0.18 NA ...
## $ b_tc               : num  NA 0 NA NA NA 0 0.67 NA 0.43 NA ..
## $ dist               : num  22.1 14 NA NA NA 17.9 16.7 NA 16.3
## $ cf                 : num  0 0 0 0 0 0 0 0 0 NA ...
## $ penm               : num  0 0 0 0 0 0 0 0 0 NA ...
## $ pent               : num  0 0 0 0 0 0 0 0 0 NA ...
## $ age                : num  24.6 25.9 19.3 32 22 ...
## - attr(*, ".internal.selfref")=<externalptr>
```

## 2.5 Relocalisation des variables

- La variable 'joueur' est 1ère position pour identifier les joueurs.
- les variables qualitatives signalitiques seront placées avant les variables continues.
- La variable 'buts' sera en 2ème position car elle sera considérée plus tard comme variable endogène dans un modèle de maching learnig avec application économétrique.

```
df= relocate(df, c("nation", "pos", "naissance"), .before = mj)
df= relocate(df, buts, .before = equipe)
colnames(df[, 1:8])
```

```
## [1] "joueur"      "buts"        "equipe"      "nation"      "pos"         "naissance"
## [7] "mj"         "min"
```

## Chapter 3

# Analyse exploratoire de la bdd

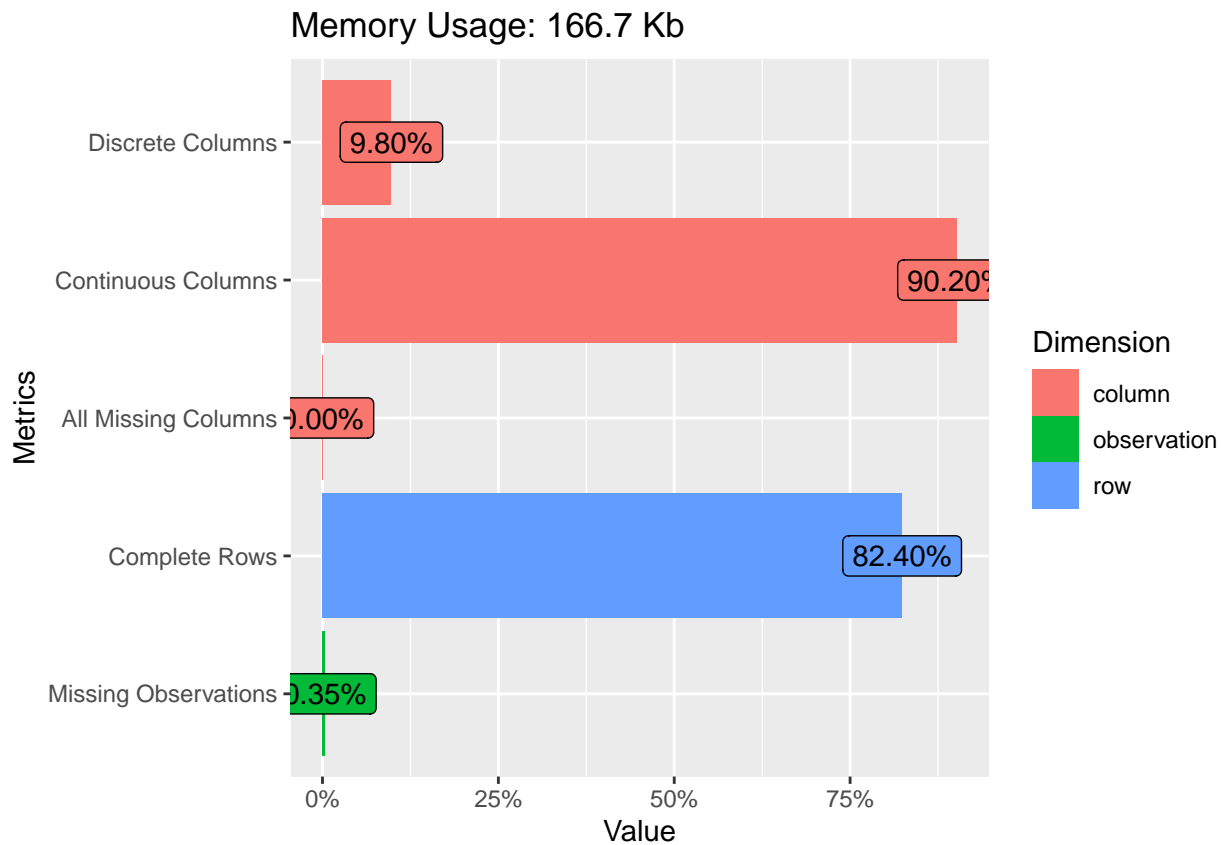
- Pour rappel, l'exploration de données dans un projet de données précède les statistiques descriptives et/ou les inférences statistiques. Cette étape est essentielle à la compréhension de la base de données en nous fournissant un premier aperçu de celle-ci.
- La variable '*buts*' ayant une grande importance dans l'analyse des performances offensives, nous ne retenons que les joueurs ayant marqué au moins 1 but dans la saison.

```
df=filter(df , buts>0)
```

Nous obtenons un total de 250 joueurs et buteurs.

### 3.1 Structure des données

```
plot_intro(df)
```



## 3.2 Données manquantes

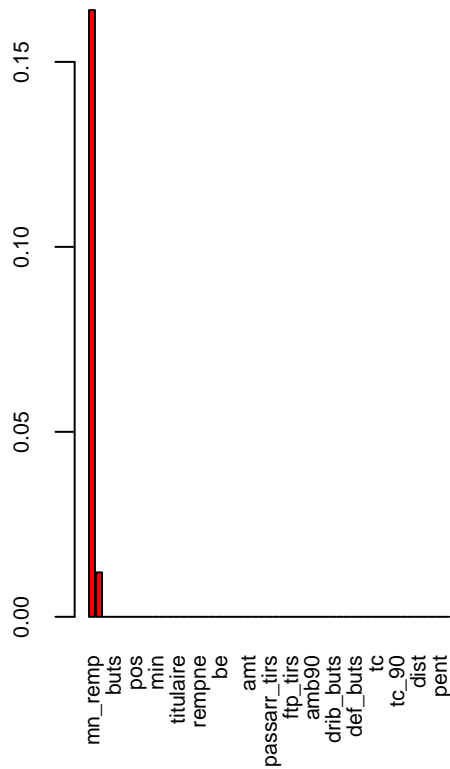
```
profile_missing(df)
```

##	feature	num_missing	pct_missing
## 1	joueur	0	0.000
## 2	but	0	0.000
## 3	equipe	0	0.000
## 4	nation	0	0.000
## 5	pos	0	0.000
## 6	naissance	0	0.000
## 7	mj	0	0.000
## 8	min	0	0.000
## 9	mn_mj	0	0.000
## 10	pourcentage_min	0	0.000
## 11	titulaire	0	0.000
## 12	mn_debute	3	0.012
## 13	compl	0	0.000
## 14	remp	0	0.000
## 15	mn_remp	41	0.164
## 16	rempne	0	0.000
## 17	ppm	0	0.000
## 18	bt	0	0.000
## 19	be	0	0.000

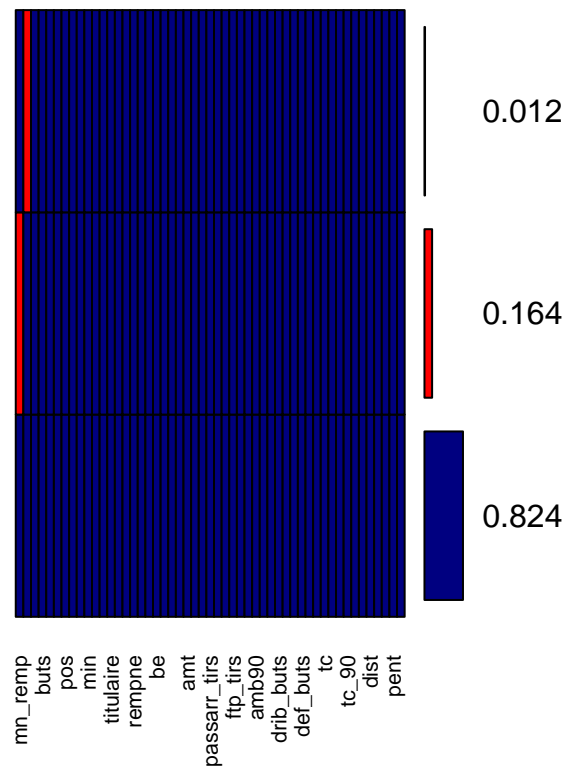
## 20	butts_marques_net_avec_joueur	0	0.000
## 21	butts_marques_net_avec_joueur_par_match	0	0.000
## 22	sur_en_dehors_du_terrain	0	0.000
## 23	amt	0	0.000
## 24	amt90	0	0.000
## 25	passjeu_tirs	0	0.000
## 26	passarr_tirs	0	0.000
## 27	drib_tirs	0	0.000
## 28	tirs_tirs	0	0.000
## 29	ftp_tirs	0	0.000
## 30	def_tirs	0	0.000
## 31	amb	0	0.000
## 32	amb90	0	0.000
## 33	passjeu_butts	0	0.000
## 34	passarr_butts	0	0.000
## 35	drib_butts	0	0.000
## 36	tirs_butts	0	0.000
## 37	ftp_butts	0	0.000
## 38	def_butts	0	0.000
## 39	minutes_jouees_90	0	0.000
## 40	tirs	0	0.000
## 41	tc	0	0.000
## 42	pourcentage_tc	0	0.000
## 43	tir_90	0	0.000
## 44	tc_90	0	0.000
## 45	b_tir	0	0.000
## 46	b_tc	0	0.000
## 47	dist	0	0.000
## 48	cf	0	0.000
## 49	penm	0	0.000
## 50	pent	0	0.000
## 51	age	0	0.000

```
df_NA <- aggr(df,
  col=c('navyblue', 'red'),
  numbers=TRUE,
  sortVars=TRUE,
  labels=names(data),
  cex.axis=.7, gap=3,
  ylab=c("Histogramme des valeurs manquantes", "Pattern"))
```

Histogramme des valeurs manquantes



Pattern



```
##
## Variables sorted by number of missings:
##      Variable Count
##      mn_remp 0.164
##      mn_debut 0.012
##      joueur 0.000
##      buts 0.000
##      equipe 0.000
##      nation 0.000
##      pos 0.000
##      naissance 0.000
##      mj 0.000
##      min 0.000
##      mn_mj 0.000
##      pourcentage_min 0.000
##      titulaire 0.000
##      compl 0.000
##      remp 0.000
##      rempne 0.000
##      ppm 0.000
##      bt 0.000
##      be 0.000
##      buts_marques_net_avec_joueur 0.000
##      buts_marques_net_avec_joueur_par_match 0.000
##      sur_en_dehors_du_terrain 0.000
##      amt 0.000
##      amt90 0.000
##      passjeu_tirs 0.000
##      passarr_tirs 0.000
```



```
##          drib_tirs 0.000
##          tirs_tirs 0.000
##          ftp_tirs 0.000
##          def_tirs 0.000
##          amb 0.000
##          amb90 0.000
##          passjeu_buts 0.000
##          passarr_buts 0.000
##          drib_buts 0.000
##          tirs_buts 0.000
##          ftp_buts 0.000
##          def_buts 0.000
##          minutes_jouees_90 0.000
##          tirs 0.000
##          tc 0.000
##          pourcentage_tc 0.000
##          tir_90 0.000
##          tc_90 0.000
##          b_tir 0.000
##          b_tc 0.000
##          dist 0.000
##          cf 0.000
##          penm 0.000
##          pent 0.000
##          age 0.000
```

```
df_NA
```

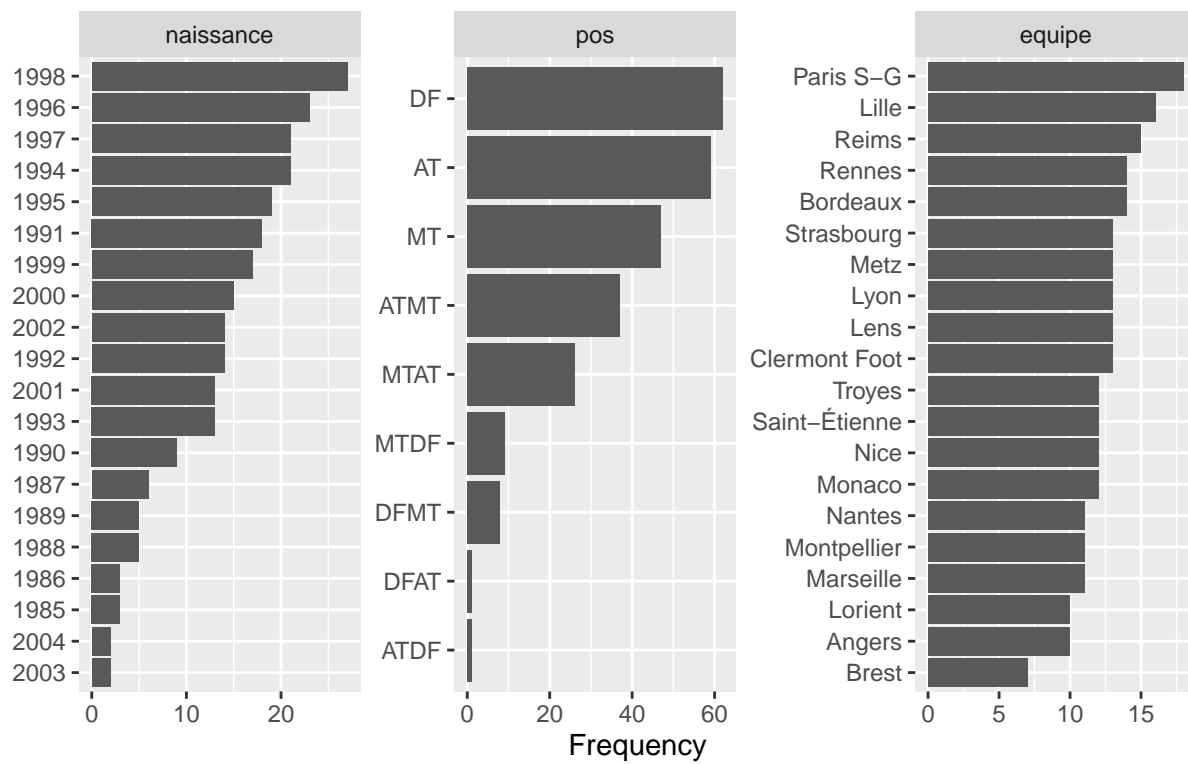
```
##
## Missings in variables:
## Variable Count
## mn_debut     3
## mn_remp      41
```

### 3.3 distribution des variables

#### 3.3.1 Variables qualitatives

```
plot_bar(
  select(df, naissance, pos, equipe),
  title = "Répartition des joueurs par caractéristiques")
```

## Répartition des joueurs par caractéristiques

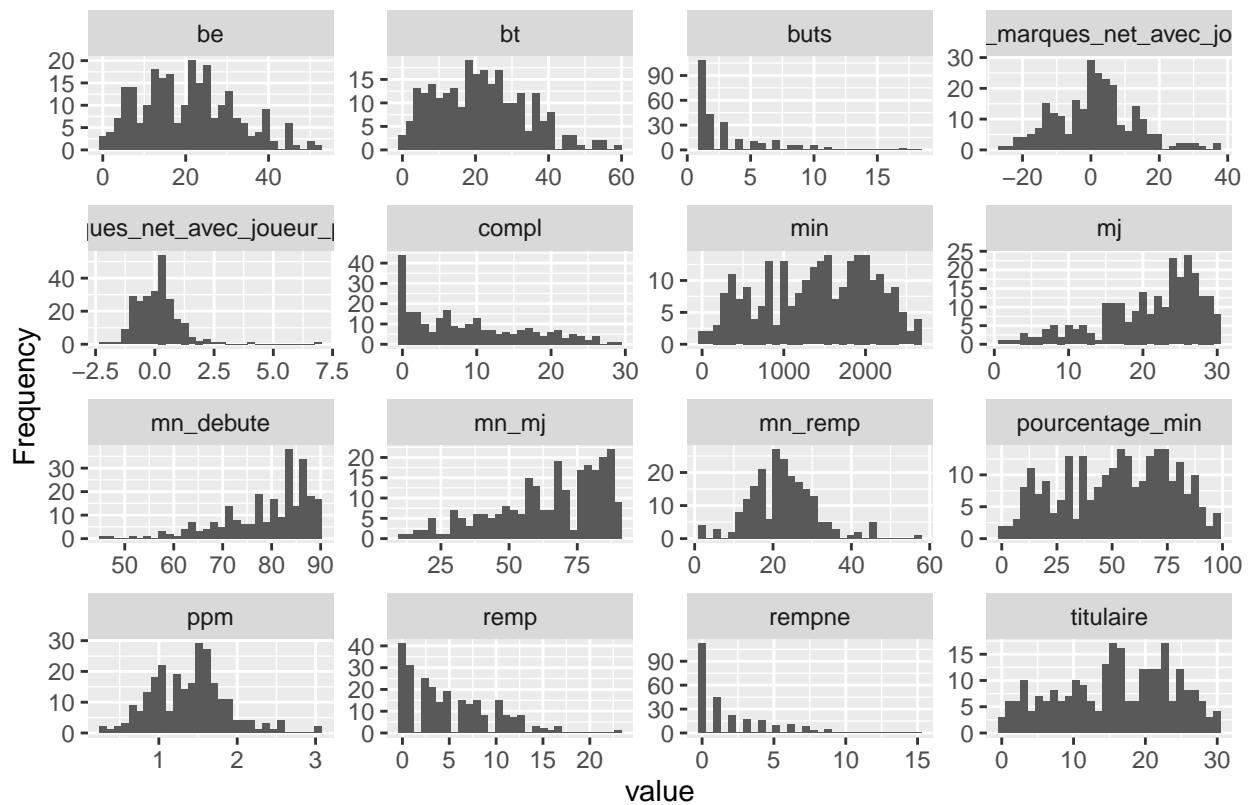


### 3.3.2 Variables continues

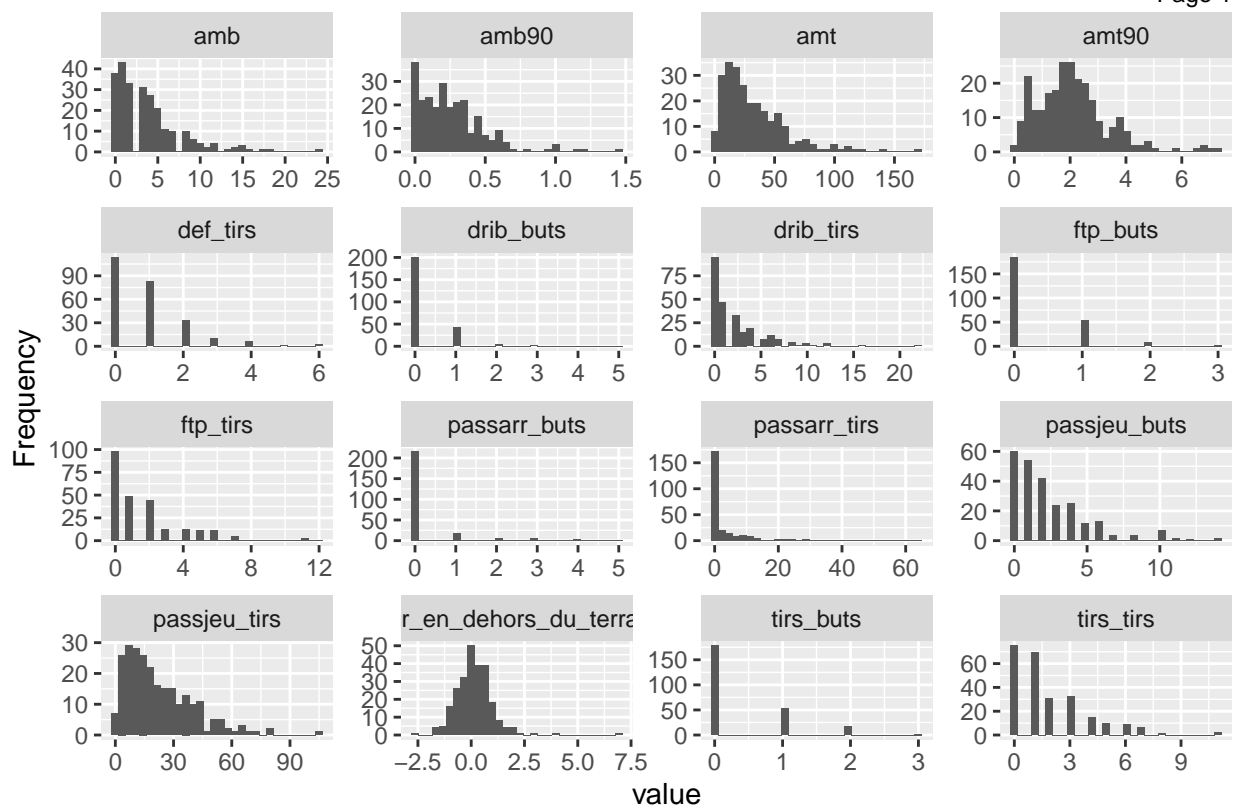
#### 3.3.2.1 Normalité

- *Histogramme*

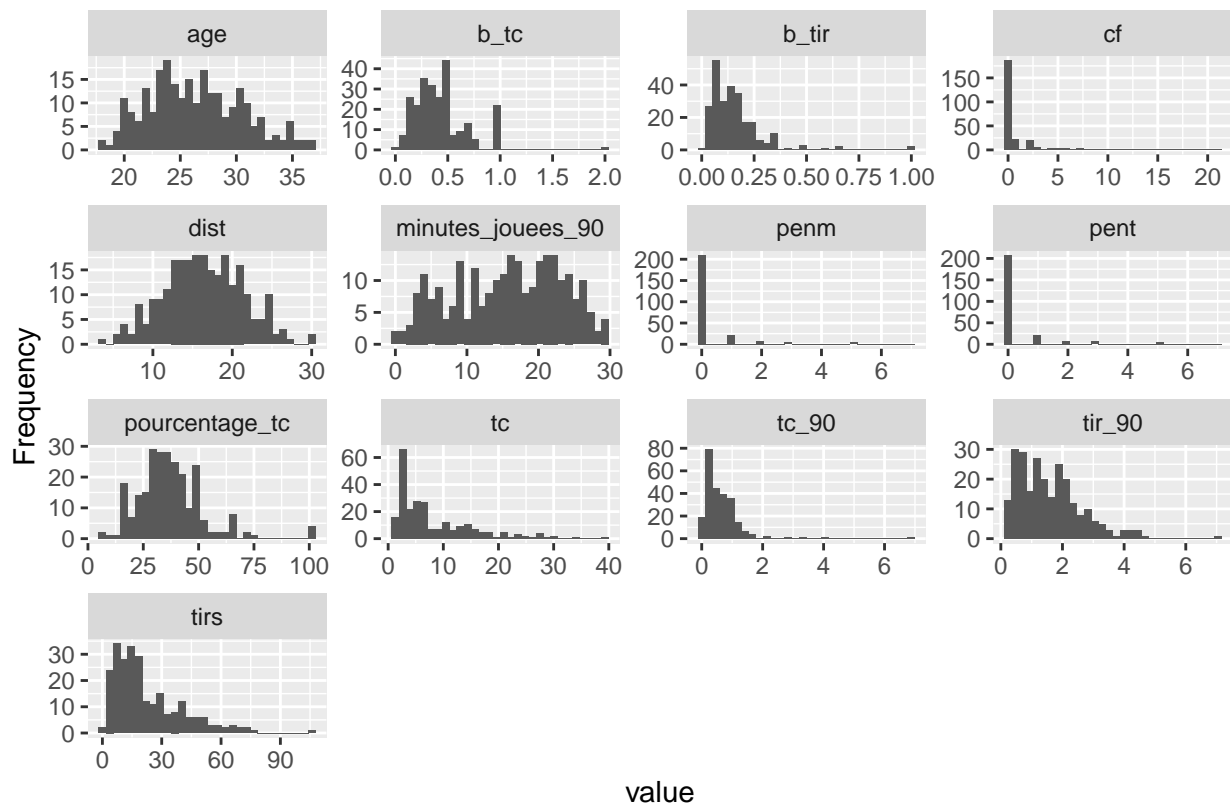
```
plot_histogram(split_columns(df)$continuous)
```



Page 1



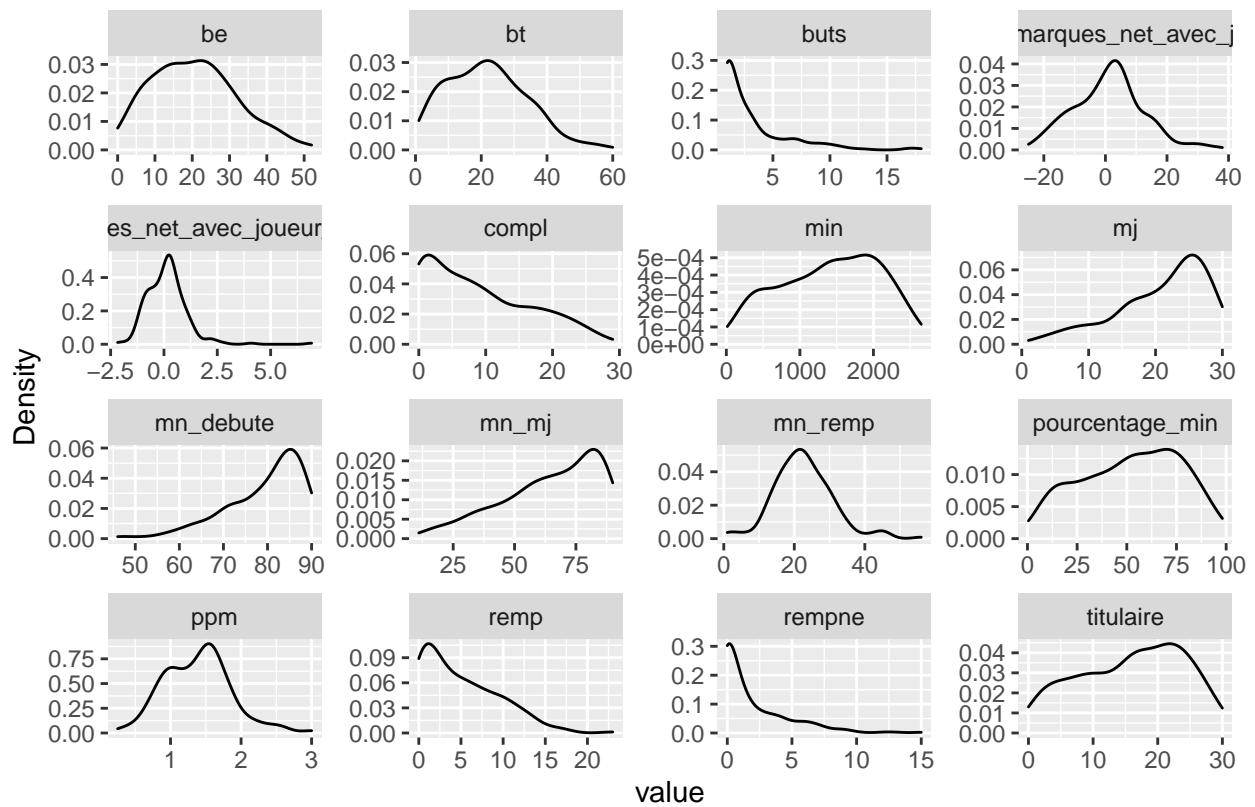
Page 2



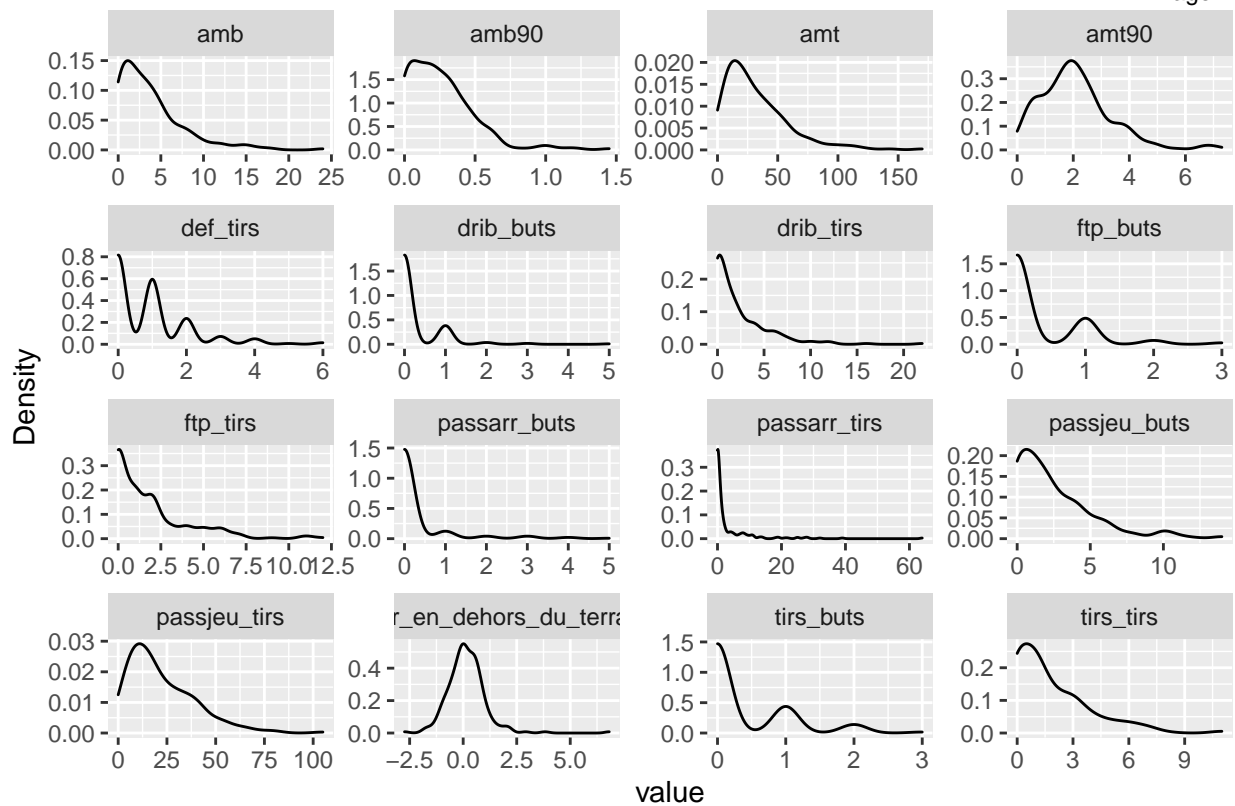
Page 3

- *Densité*

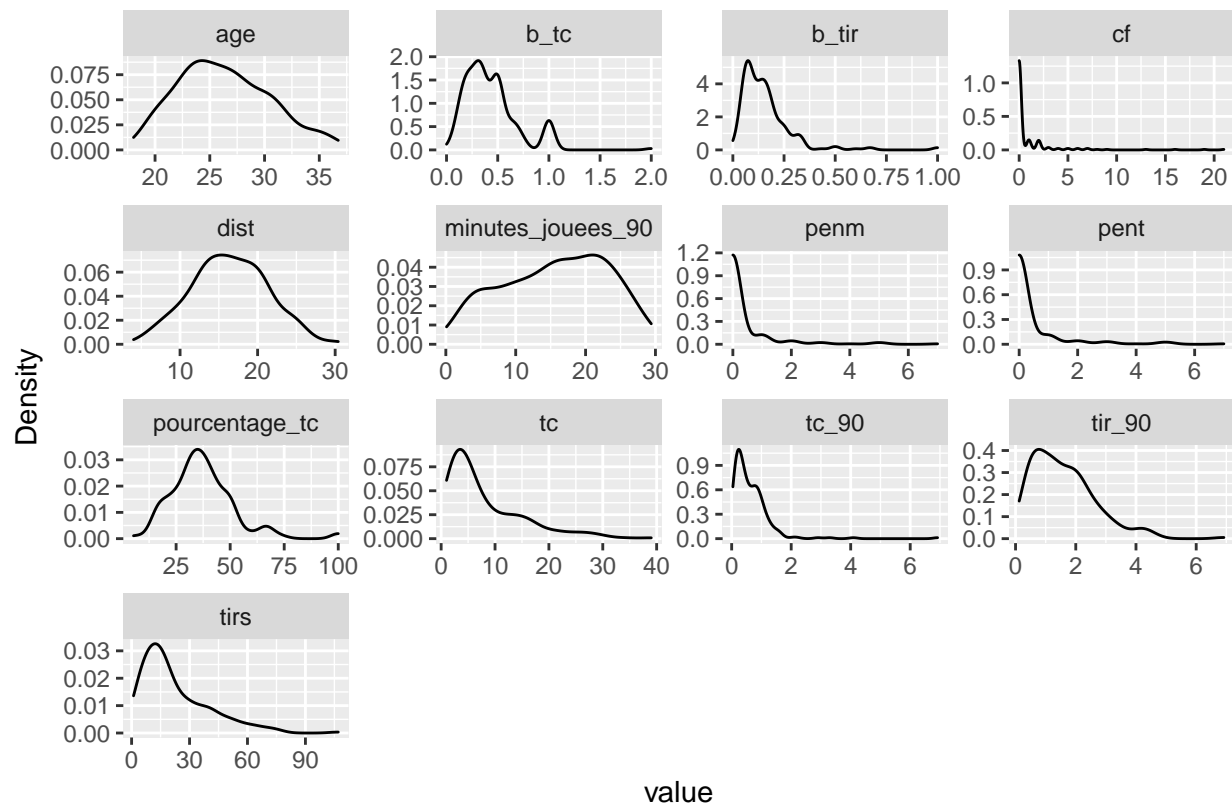
```
plot_density(split_columns(df)$continuous)
```



Page 1



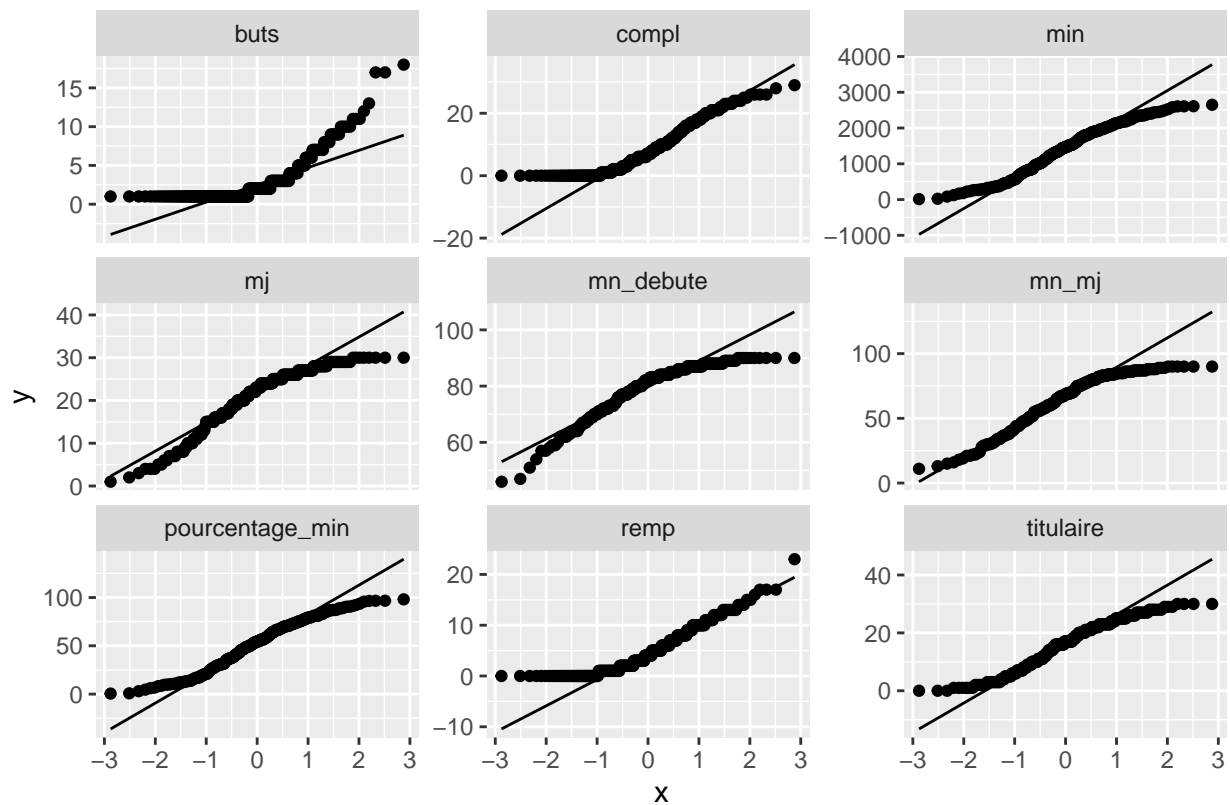
Page 2



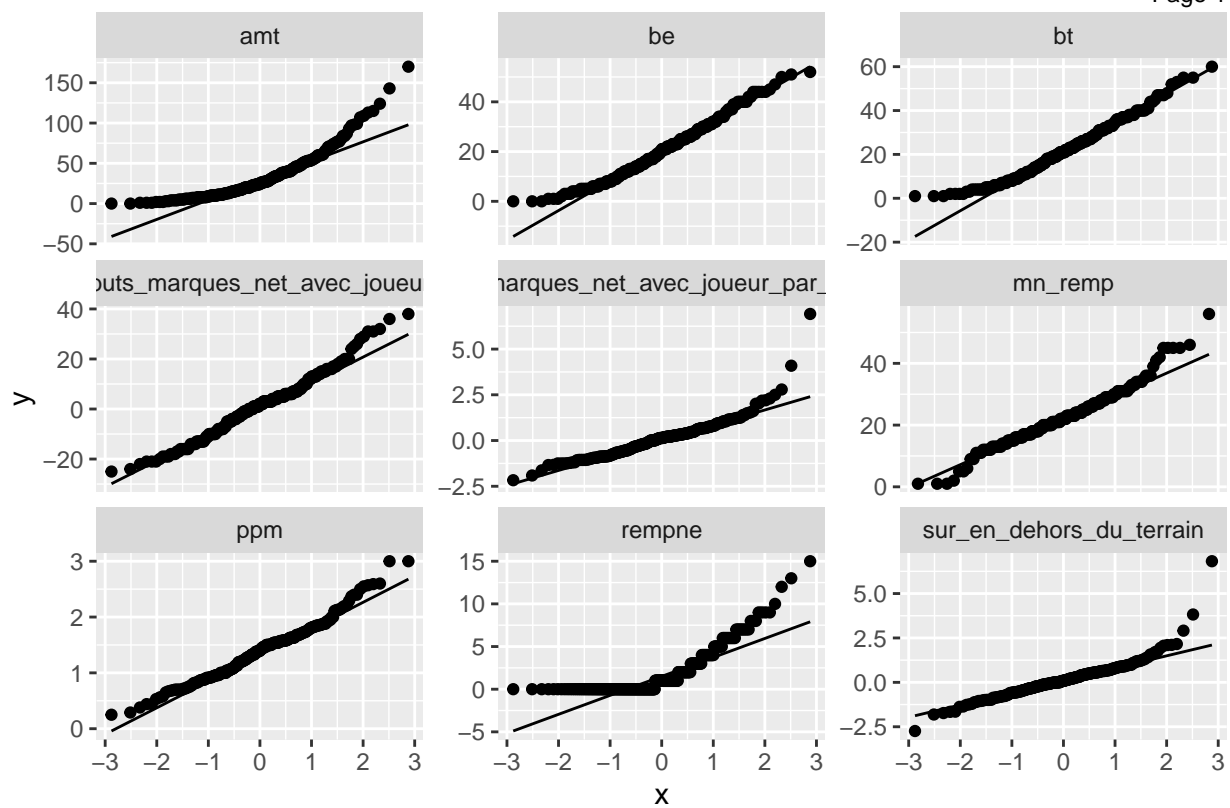
Page 3

- *qq-plot*

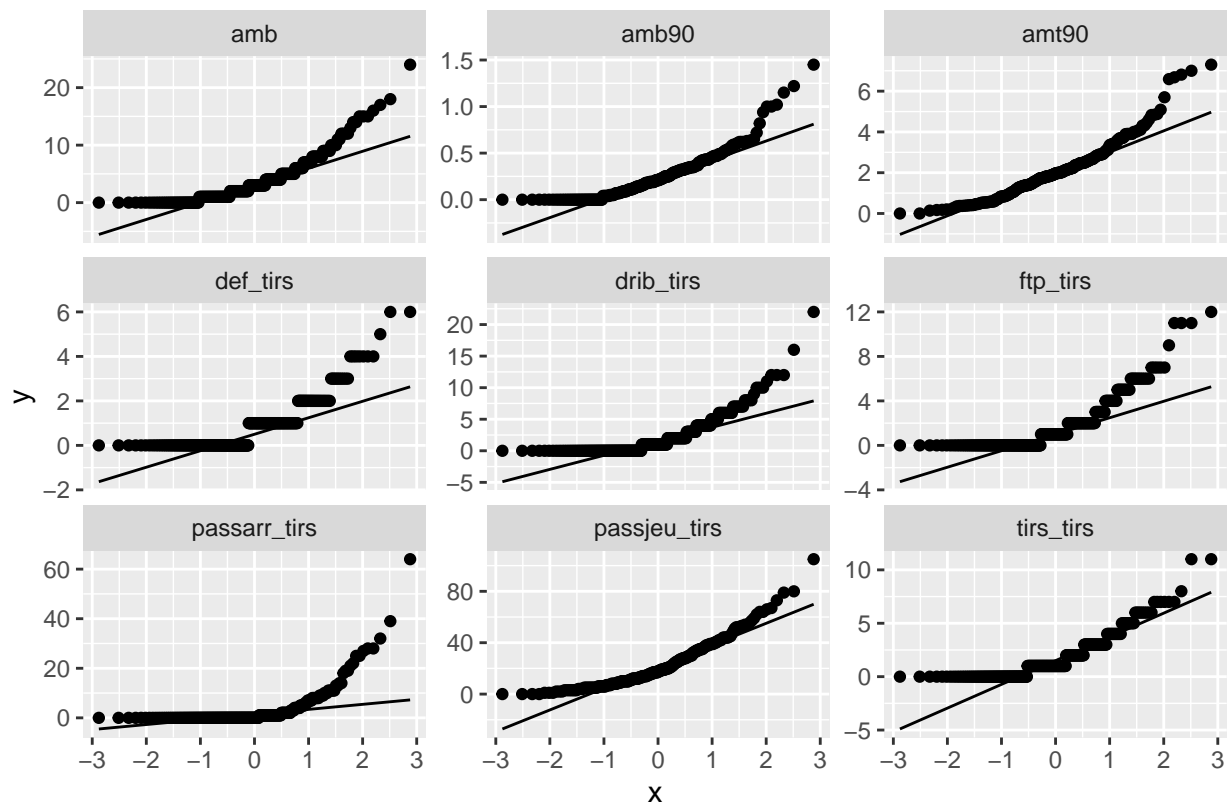
```
plot_qq(split_columns(df)$continuous)
```



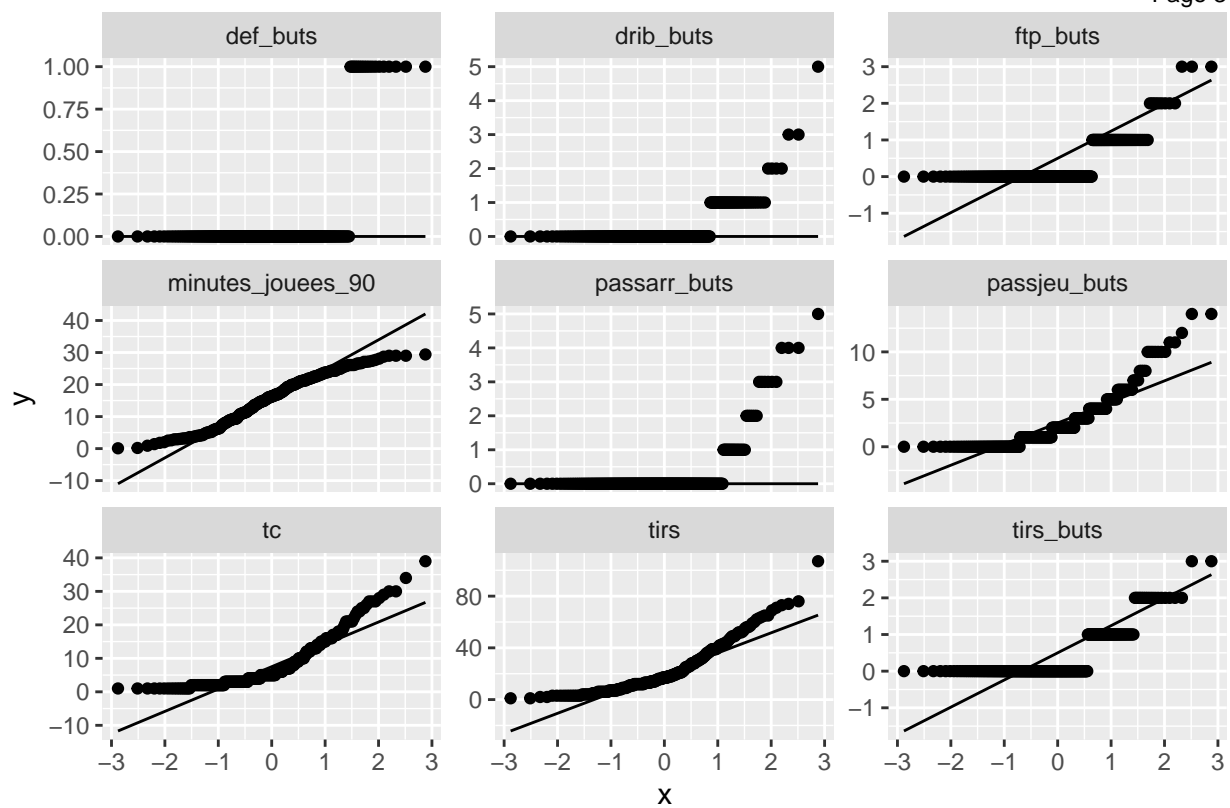
Page 1



Page 2

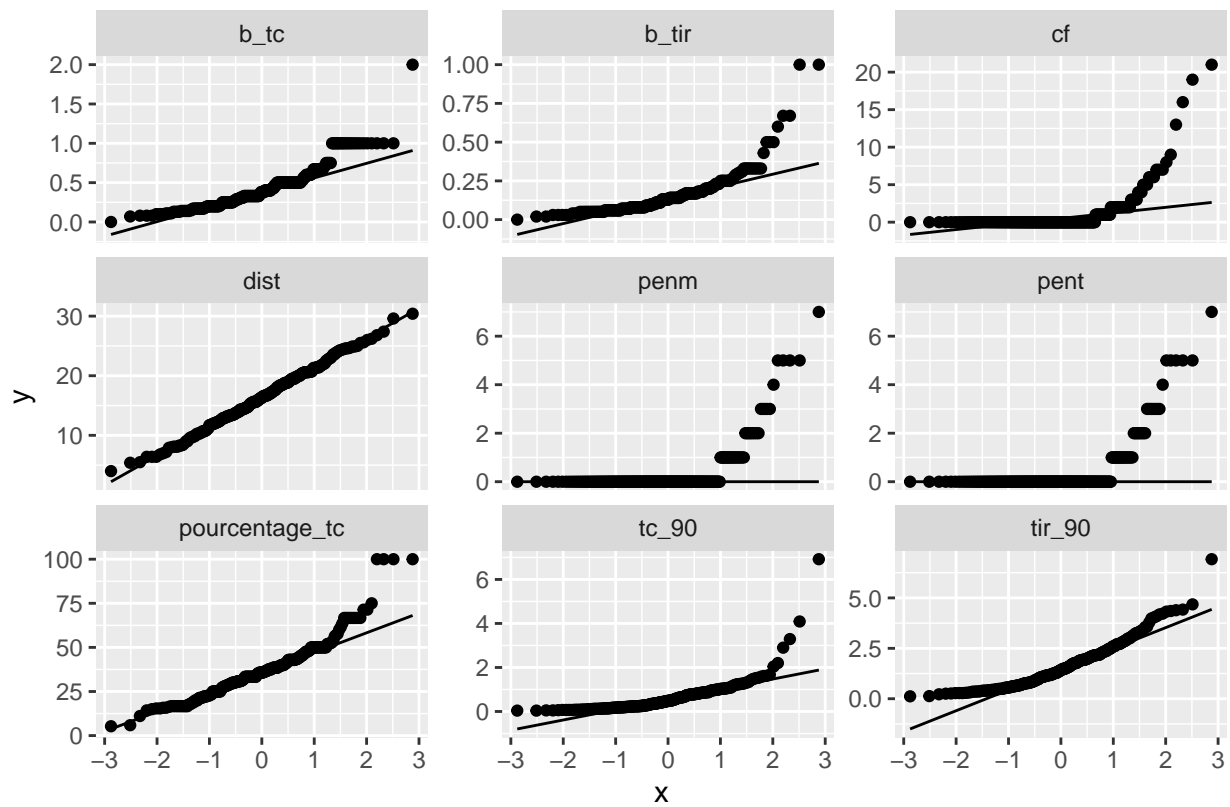


Page 3

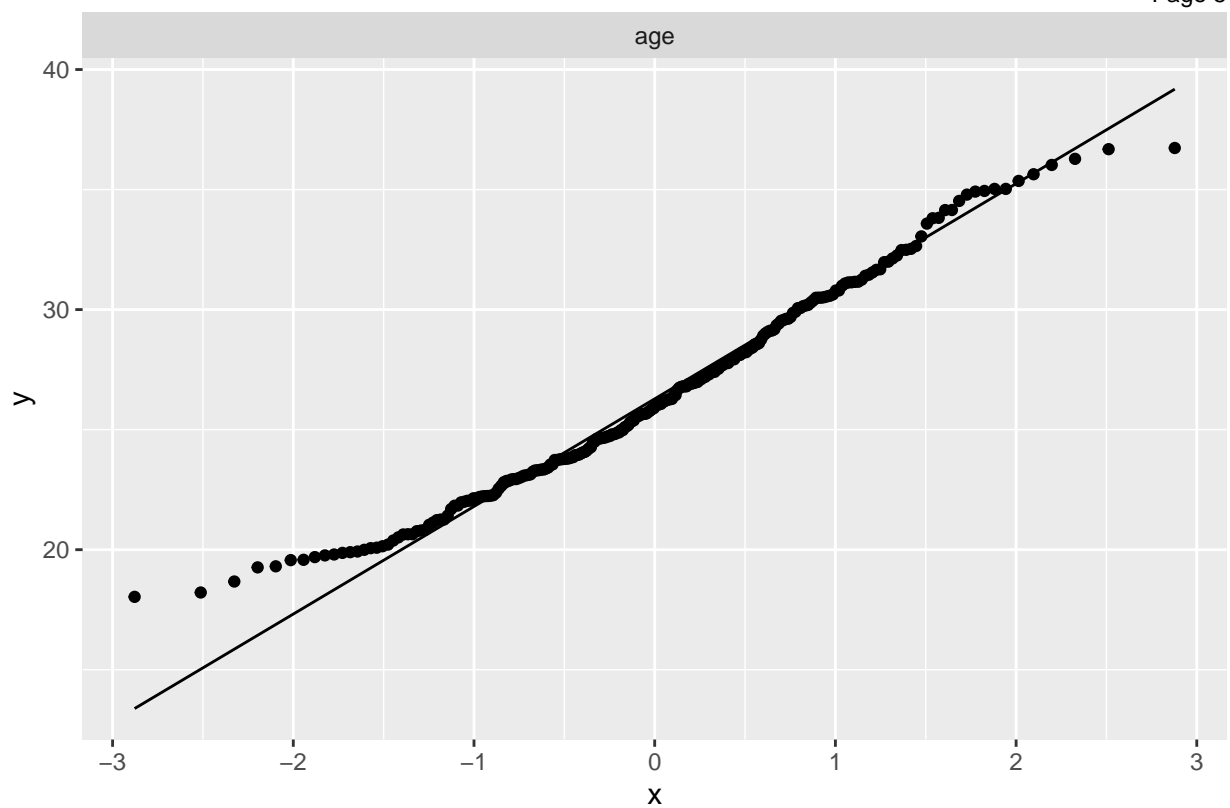


Page 4





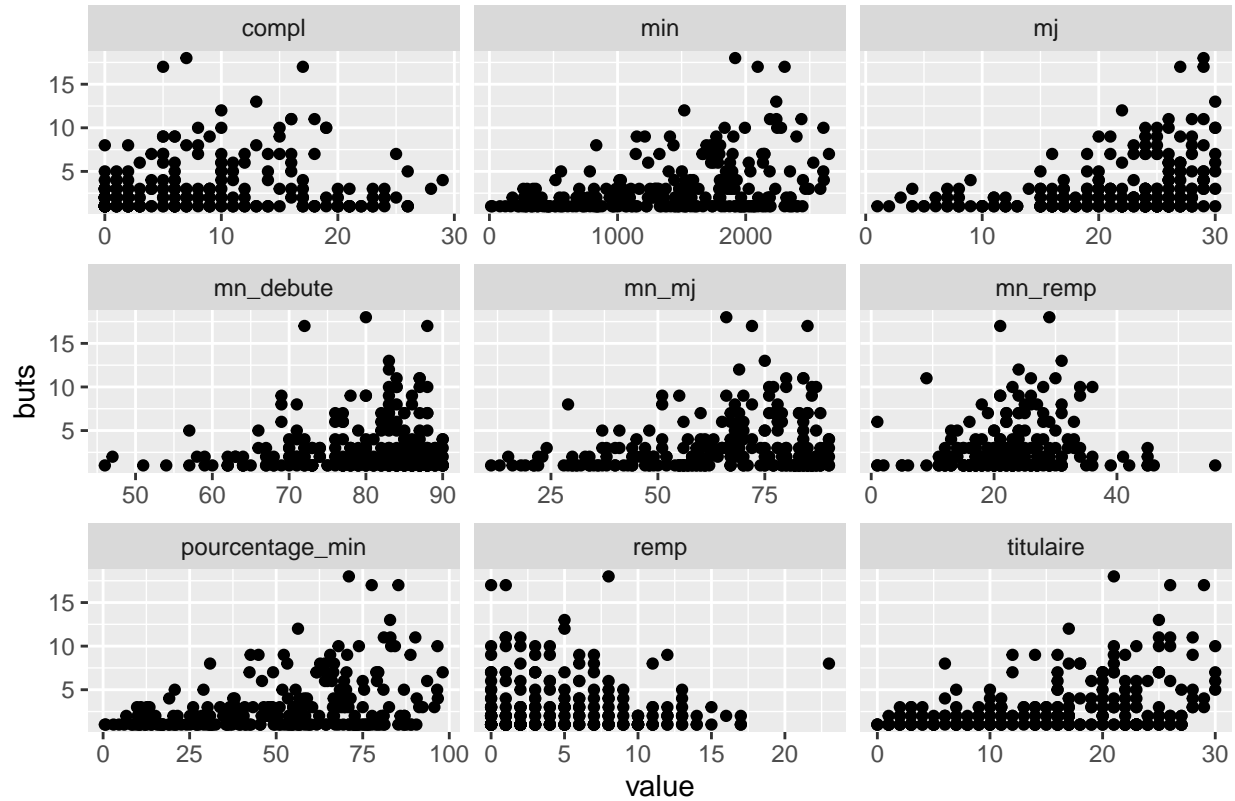
Page 5



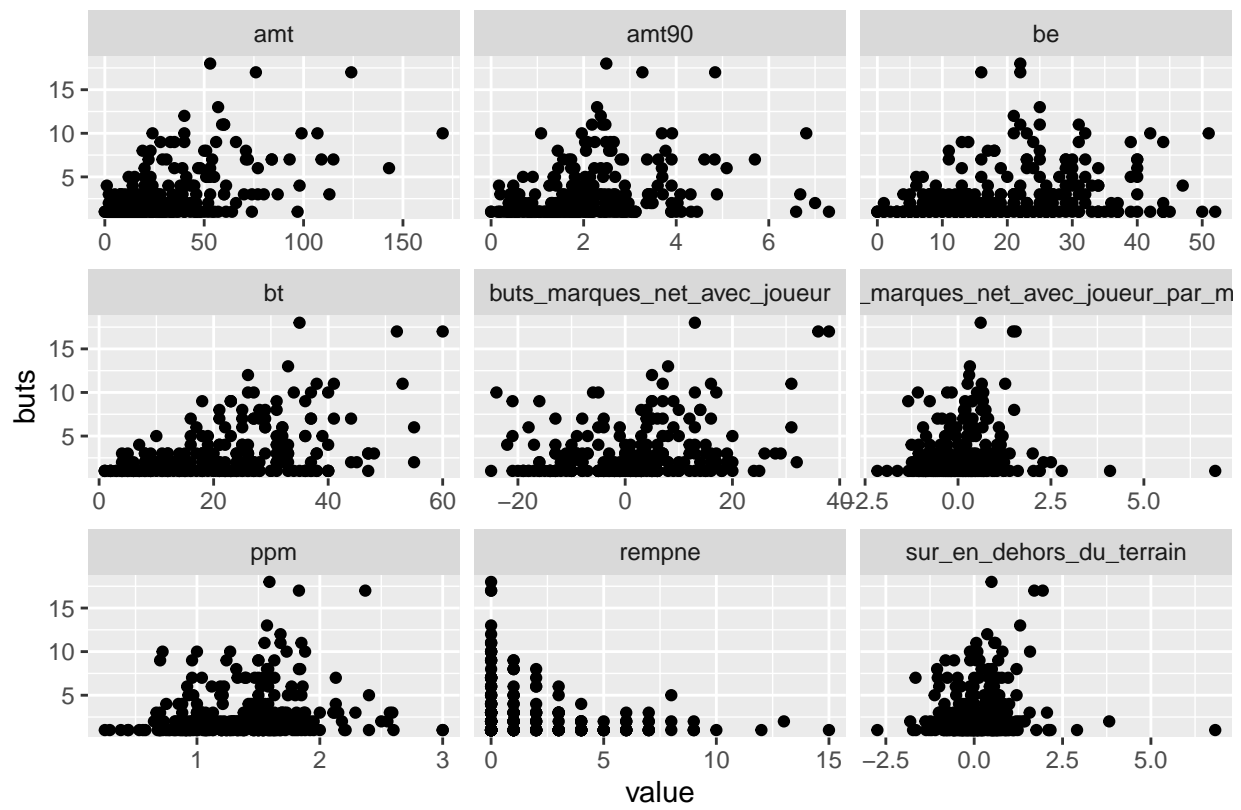
Page 6

### 3.3.2.2 Evolution du nombre de buts en fonction des autres variables continues

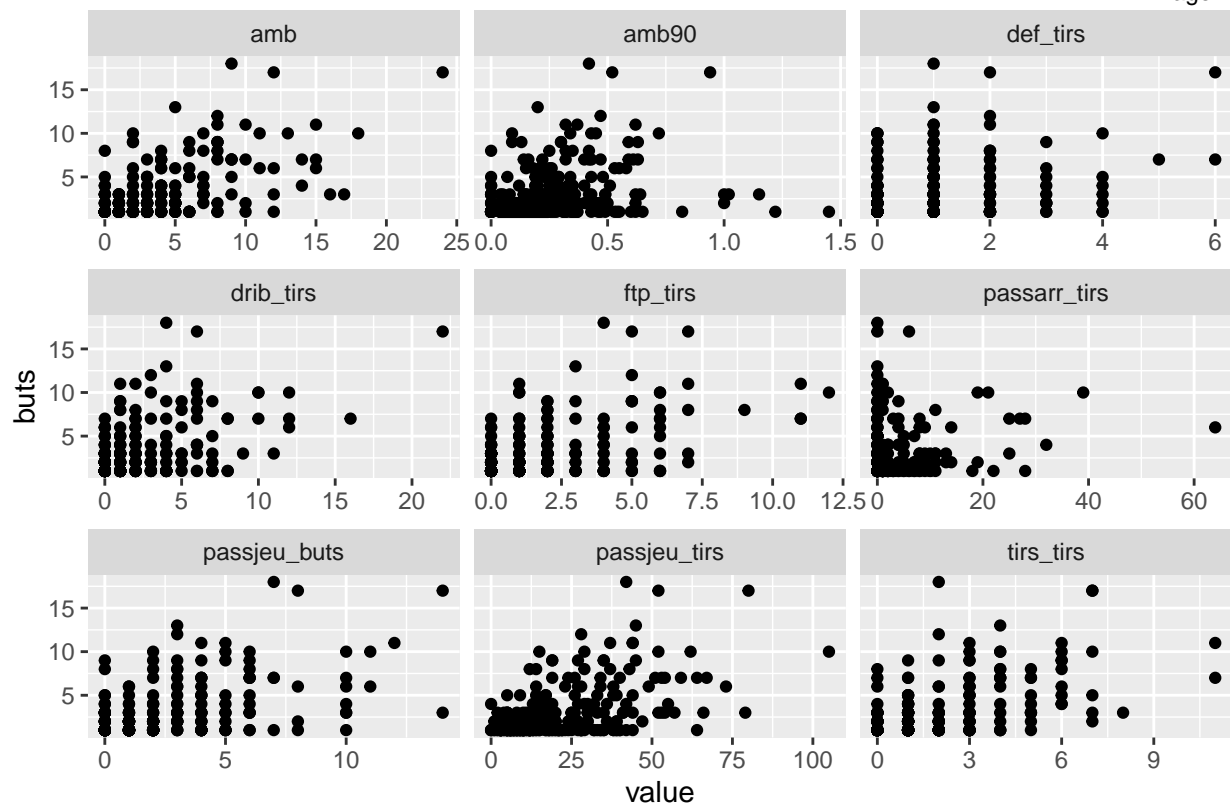
```
plot_scatterplot(split_columns(df)$continuous, by = "buts")
```



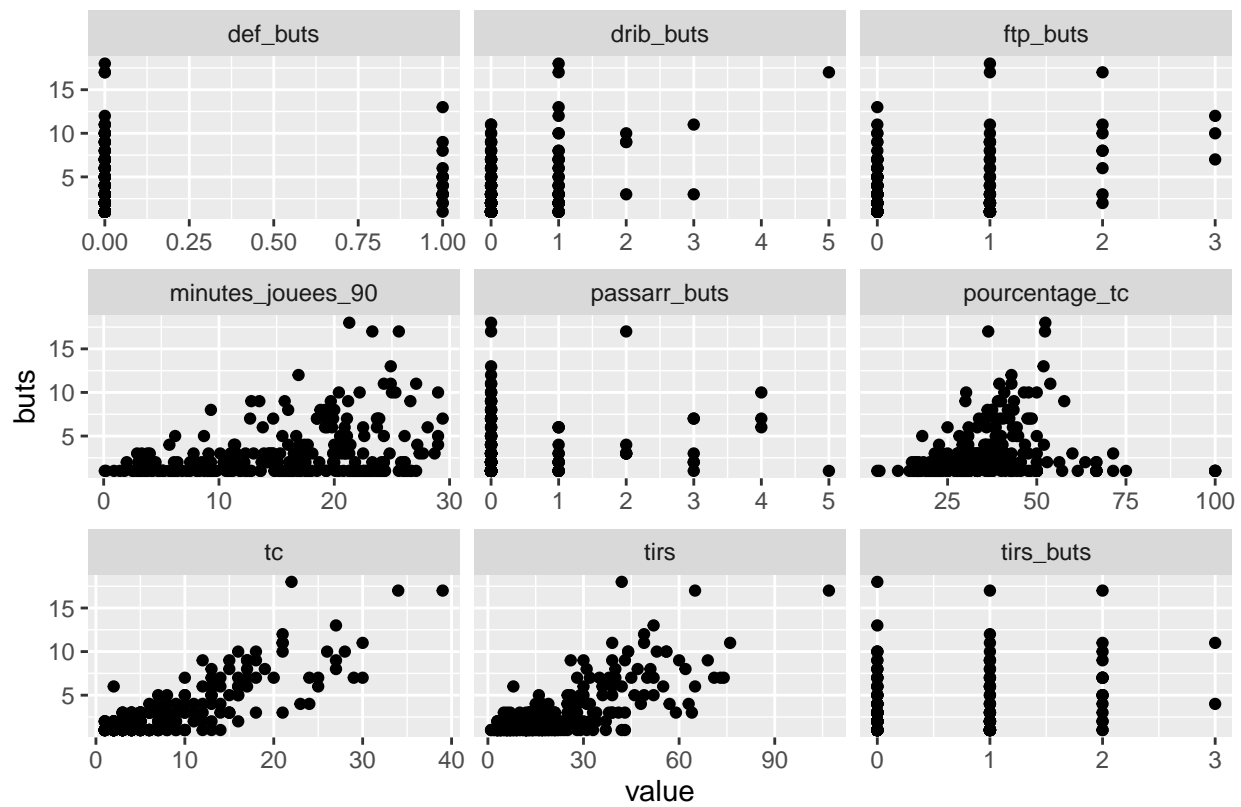
Page 1



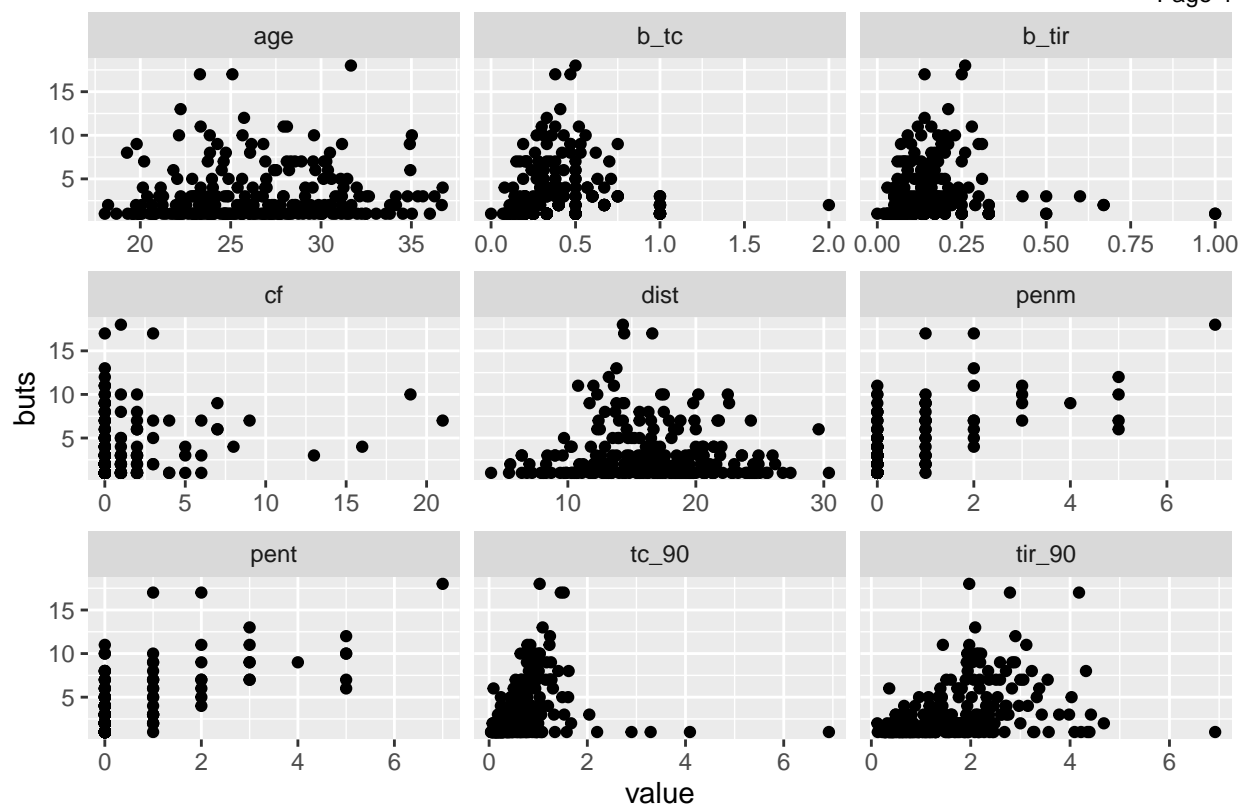
Page 2



Page 3



Page 4



Page 5

Les variables 'def\_tirs', 'ftp\_tirs', 'passjeu\_buts', 'tirs\_tirs', 'def\_buts,drib\_buts', 'ftp\_buts','passarr\_buts', 'tirs\_buts',

'penm' et 'pent' seront discrétisés.

```
df_a_discritise = select(df,
                          def_tirs,ftp_tirs,passjeu_buts,
                          tirs_tirs,def_buts,drib_buts,
                          ftp_buts,passarr_buts,tirs_buts,
                          penm,pent)

df_a_discritise = update_columns(
  df_a_discritise,
  c("def_tirs","ftp_tirs","passjeu_buts","tirs_tirs",
    "def_buts","drib_buts","ftp_buts","passarr_buts",
    "tirs_buts","penm","pent")
    , as.factor)
colnames(df_a_discritise) <- paste("discr",
                                   colnames(df_a_discritise),
                                   sep="_")

summary(df_a_discritise)
```

```
##  discr_def_tirs  discr_ftp_tirs  discr_passjeu_buts  discr_tirs_tirs
##  0:114          0      :98      0      :60          0      :75
##  1: 83          1      :49      1      :54          1      :69
##  2: 33          2      :45      2      :42          3      :32
##  3: 10          3      :13      4      :25          2      :31
##  4: 7           4      :13      3      :24          4      :15
##  5: 1           5      :11      6      :13          5      :10
##  6: 2           (Other):21      (Other):32          (Other):18
##  discr_def_buts  discr_drib_buts  discr_ftp_buts  discr_passarr_buts
##  0:232          0:201          0:185          0:216
##  1: 18          1: 42          1: 54          1: 18
##                2: 4          2: 8          2: 6
##                3: 2          3: 3          3: 6
##                5: 1          4: 3
##                5: 1
##
##  discr_tirs_buts  discr_penm  discr_pent
##  0:178          0:210      0:208
##  1: 53          1: 22      1: 21
##  2: 17          2: 8       2: 8
##  3: 2           3: 4       3: 6
##                4: 1       4: 1
##                5: 4       5: 5
##                7: 1       7: 1
```

```
df1=data.frame(df,df_a_discritise)
```

- Nous obtenons ainsi une nouvelle base de donnée finale 'df1' de dimension (250, 62) avec les variables discrétisées en plus (contre 'df' de dimension (250, 51)). Cette base 'nettoyée' servira ainsi à réaliser une analyse statistique par la suite.
- En outre, le raffinement de cette purification de la donnée est susceptible d'être complété au cours de la suite de l'analyse en fonction de l'évolution de la compréhension de l'étude tant d'un point de vue qualitatif que quantitatif.