

TD n° 4 : Nettoyage des données

Exercice 1. On travaille sur le jeu de données `airquality` inclus dans le package `datasets` (déjà chargé par défaut dans R).

1. Indiquer la présence de valeurs manquantes dans le jeu de données.
2. Créer un jeu de données dépourvu de données manquantes.
3. Proposer une analyse statistique globale (résumé) des données.

Exercice 2. On travaille avec le jeu données artificiel donné par les commandes R suivantes :

```
donnees = data.frame(x1 = c(1:4, 71617, 1, NA, 1, 1, NA),
x1 = c(1:5, 1, "NA", 1, 1, "NA"),
x1 = c(letters[c(3:5)], "x x", "x", " y y", "x", "a", "a", NA),
x4 = "",
x5 = NA,
x6 = c("128.67", "282,78", "272 eur", "262,2", "they 272", "289.2", "172 eur",
"128$", "128$", "money 272"))
donnees
```

Cela renvoie :

	x1	x1.1		x1.2	x4	x5		x6
1	1	1		c		NA		128.67
2	2	2		d		NA		282,78
3	3	3		e		NA		272 eur
4	4	4		x	x	NA		262,2
5	71617	5		x		NA	they	272
6	1	1	y	y	y	NA		289.2
7	NA	NA		x		NA		172 eur
8	1	1		a		NA		128\$
9	1	1		a		NA		128\$
10	NA	NA		<NA>		NA	money	272

Dès lors, reproduire et comprendre l'enjeu des commandes suivantes :

```
colnames(donnees) = paste0("col", 1:ncol(donnees))
donnees
```

```
donnees[donnees == ""]
donnees[donnees == ""] = NA
donnees
```

```
class(donnees$col2)
```

```
donnees$col2[donnees$col2 == "NA"] = NA
donnees

donnees = donnees[rowSums(is.na(donnees)) != ncol(donnees), ]
donnees

donnees = donnees[ , colSums(is.na(donnees)) != nrow(donnees)]
donnees

donnees = na.omit(donnees)
donnees

donnees = unique(donnees)
donnees

library(cleaner) # utiliser que pour la commande suivante
donnees$col6 = clean_numeric(donnees$col6)

sapply(donnees, class)

donnees = type.convert(donnees, as.is = TRUE)
donnees

sapply(donnees, class)

donnees$col1[donnees$col1 %in% boxplot.stats(donnees$col1)$out]

donnees = donnees[! donnees$col1 %in% boxplot.stats(donnees$col1)$out, ]
donnees

donnees$col3 = gsub(" ", "", donnees$col3)
donnees
```

On doit alors aboutir à une version nettoyée des données, à savoir:

	col1	col2	col3	col6
1	1	1	c	128.67
2	2	2	d	282.78
3	3	3	e	272.00
4	4	4	xx	262.20
6	1	1	yyy	289.20
8	1	1	a	128.00