Projet d'étude de Statistiques

Maxime Baba, Alexandre Demarquet, Félix de Brandois, Tristan Gay

2023-11-09

Contents

0.1 R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(data)
```

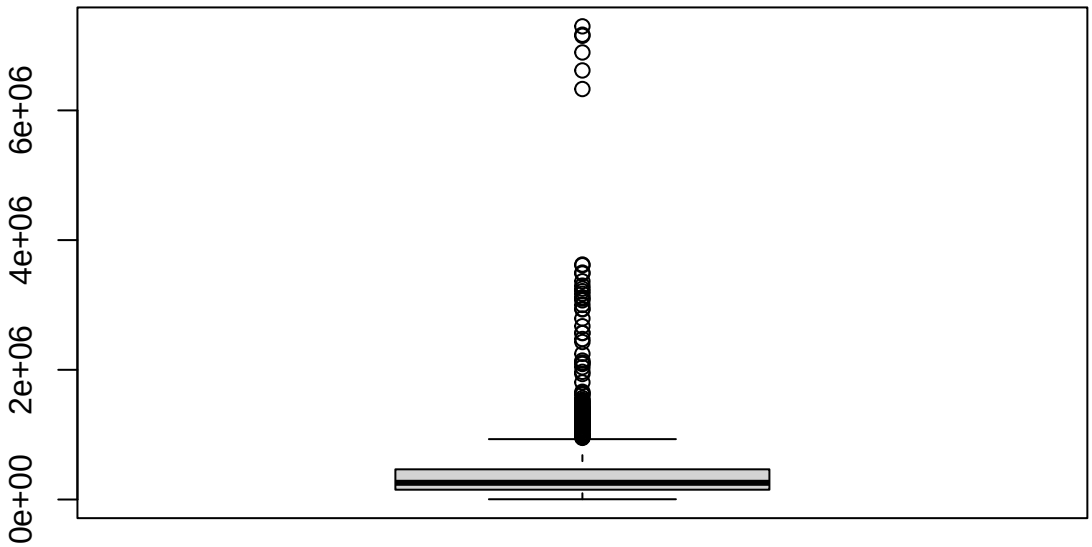
```
##    code_epci           lib_epci           annee_inv        nox_kg
##  Min.   :200006930   Length:984         Min.   :2014   Min.   :   4085
##  1st Qu.:200066342   Class :character   1st Qu.:2015   1st Qu.: 151025
##  Median :200071178   Mode  :character   Median :2016   Median : 257531
##  Mean   :219957739                      Mean   :2016   Mean   : 466703
##  3rd Qu.:243200433                      3rd Qu.:2018   3rd Qu.: 465286
##  Max.   :248400251                      Max.   :2019   Max.   :7296028
##      so2_kg            pm10_kg           pm25_kg           co_kg
##  Min.   :    91.3   Min.   :   2366   Min.   :  1971   Min.   :   28231
##  1st Qu.:  3694.6   1st Qu.:  56986   1st Qu.: 40502   1st Qu.:  482779
##  Median :  6302.6   Median : 104200   Median : 65338   Median :  764707
##  Mean   : 21041.6   Mean   : 142992   Mean   : 93646   Mean   : 1117926
##  3rd Qu.: 11327.3   3rd Qu.: 172210   3rd Qu.:106391   3rd Qu.: 1169215
##  Max.   :713262.5   Max.   :1186604   Max.   :927205   Max.   :12302750
##     c6h6_kg             nh3_kg          ges_teqco2          ch4_t
##  Min.   :   437.2   Min.   :   7408   Min.   :   2481   Min.   :  30.57
##  1st Qu.:  8497.0   1st Qu.:  83409   1st Qu.:  76282   1st Qu.: 276.97
##  Median : 12599.3   Median : 216921   Median : 118249   Median : 533.78
##  Mean   : 17863.6   Mean   : 263843   Mean   : 209643   Mean   : 905.64
##  3rd Qu.: 19764.0   3rd Qu.: 364883   3rd Qu.: 212410   3rd Qu.:1074.44
##  Max.   :178741.5   Max.   :1202364   Max.   :3171595   Max.   :5313.16
##      co2_t               n2o_t          TypeEPCI          nomdepart
##  Min.   :    865.1   Min.   :  0.952   Length:984         Length:984
##  1st Qu.:  31906.3   1st Qu.: 18.017   Class :character   Class :character
##  Median :  59008.8   Median : 41.718   Mode  :character   Mode  :character
##  Mean   : 138412.7   Mean   : 48.855
```

```
##  3rd Qu.: 131054.4   3rd Qu.: 65.378
##  Max.   :2678898.2   Max.   :726.911
##     Ardèche             Ariège              Aude             Aveyron
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.000000   Median :0.00000   Median :0.00000   Median :0.000
##  Mean   :0.006098   Mean   :0.04878   Mean   :0.06098   Mean   :0.122
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000
##       Gard        Haute.Garonne          Gers            Hérault
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1037   Mean   :0.1159   Mean   :0.1037   Mean   :0.1037
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      Landes              Lot          Lot.et.Garonne         Lozère
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.000000   Median :0.00000   Median :0.000000   Median :0.00000
##  Mean   :0.006098   Mean   :0.06098   Mean   :0.006098   Mean   :0.06707
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000
##  Pyrénées.Atlantiques Hautes.Pyrénées   Pyrénées.Orientales      Tarn
##  Min.   :0.0000       Min.   :0.00000   Min.   :0.00000     Min.   :0.00000
##  1st Qu.:0.0000       1st Qu.:0.00000   1st Qu.:0.00000     1st Qu.:0.00000
##  Median :0.0000       Median :0.00000   Median :0.00000     Median :0.00000
##  Mean   :0.0122       Mean   :0.06098   Mean   :0.07317     Mean   :0.09756
##  3rd Qu.:0.0000       3rd Qu.:0.00000   3rd Qu.:0.00000     3rd Qu.:0.00000
##  Max.   :1.0000       Max.   :1.00000   Max.   :1.00000     Max.   :1.00000
##  Tarn.et.Garonne      Vaucluse            latit           longit
##  Min.   :0.00000   Min.   :0.000000   Min.   :42.44   Min.   :-0.295
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:43.33   1st Qu.: 1.257
##  Median :0.00000   Median :0.000000   Median :43.70   Median : 2.157
##  Mean   :0.06098   Mean   :0.006098   Mean   :43.70   Mean   : 2.159
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:44.10   3rd Qu.: 3.034
##  Max.   :1.00000   Max.   :1.000000   Max.   :44.88   Max.   : 4.825
```

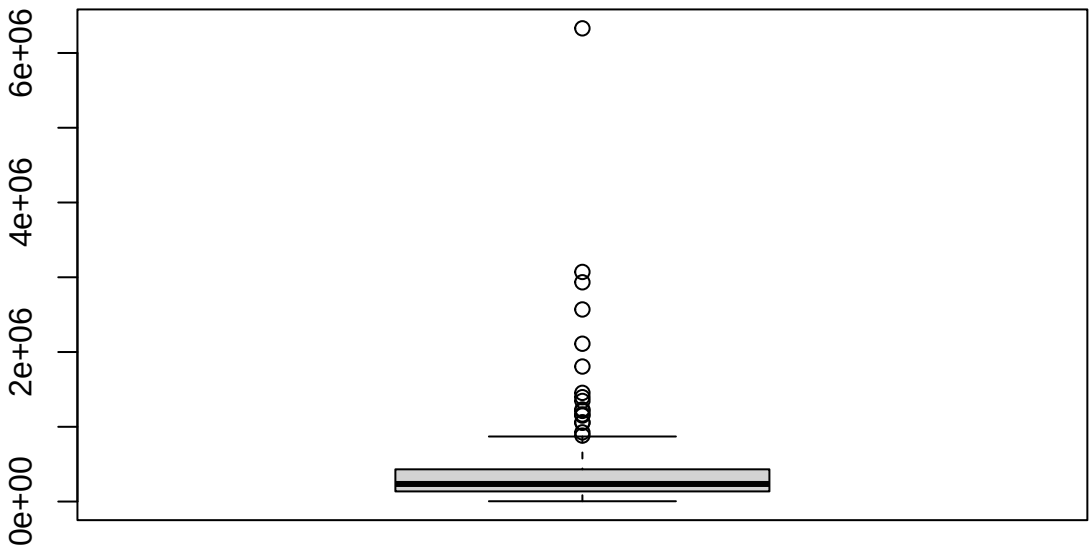
```
data[1,]
```

```
##   code_epci          lib_epci annee_inv   nox_kg   so2_kg  pm10_kg  pm25_kg
## 1 200006930 CC du Haut Allier      2019 65633.66 3866.599 15728.87 10975.55
##      co_kg  c6h6_kg   nh3_kg ges_teqco2   ch4_t    co2_t  n2o_t TypeEPCI
## 1 173194.3 2319.199 133686.2   43995.12 617.104 17831.59 17.114       CC
##   nomdepart Ardèche Ariège Aude Aveyron Gard Haute.Garonne Gers Hérault Landes
## 1    Lozère       0      0    0       0    0             0    0       0      0
##   Lot Lot.et.Garonne Lozère Pyrénées.Atlantiques Hautes.Pyrénées
## 1   0              0      1                    0               0
##   Pyrénées.Orientales Tarn Tarn.et.Garonne Vaucluse   latit   longit
## 1                   0    0               0        0 44.7324 3.769267
```

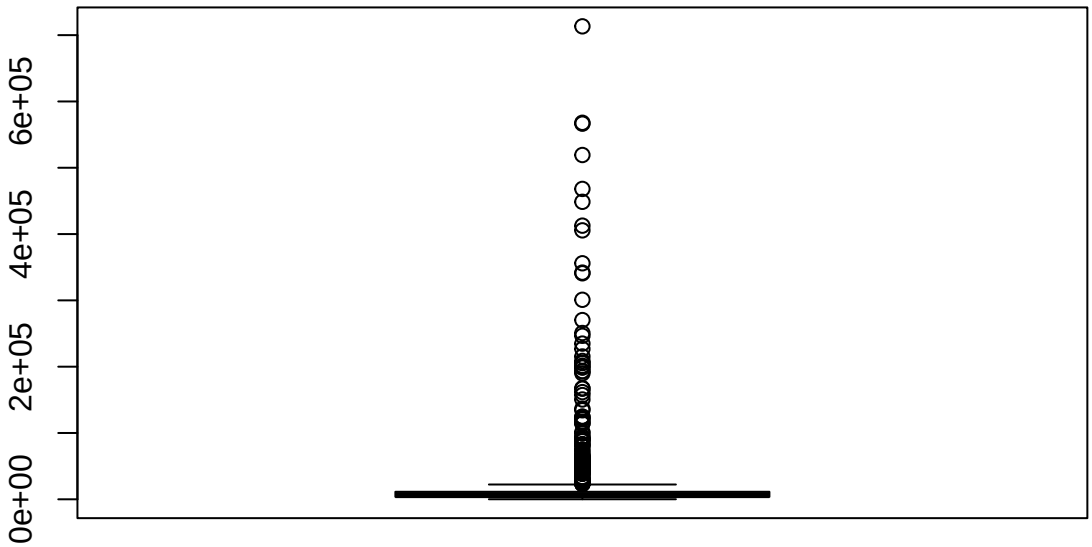
```
boxplot(data$nox_kg)
```



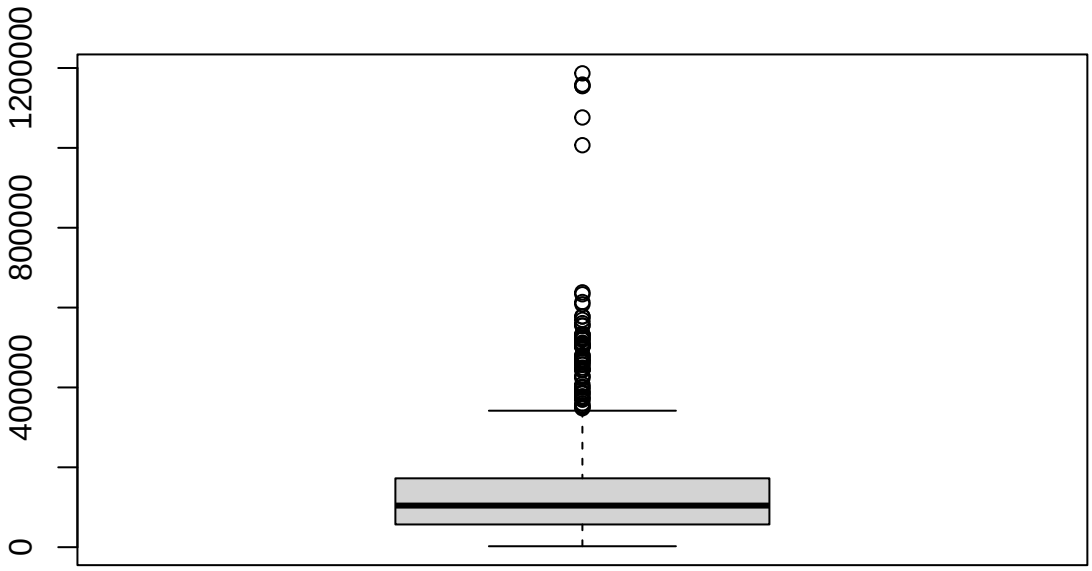

```
boxplot(data$nox_kg[1:164])
```



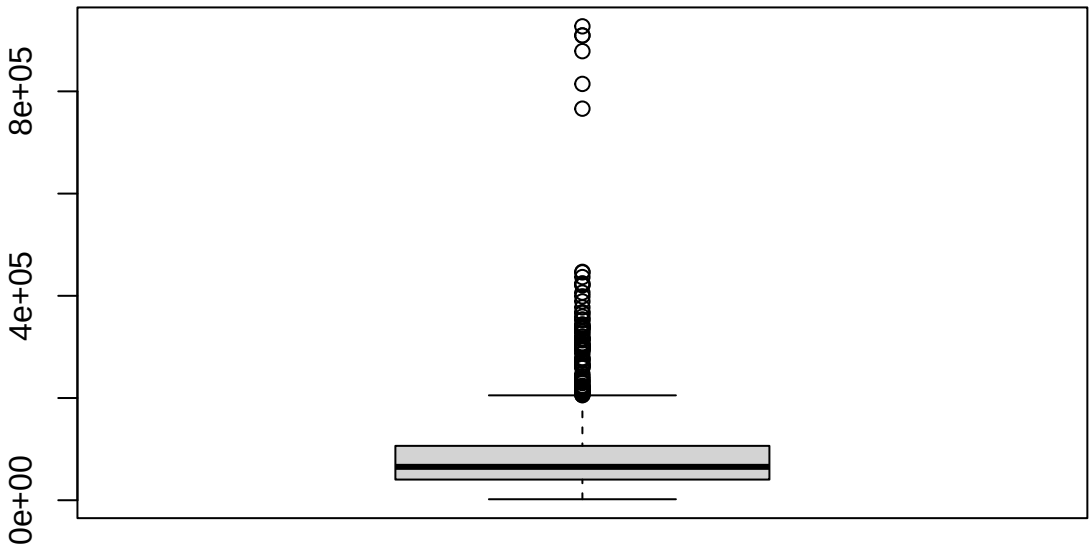

```
boxplot(data$so2_kg)
```



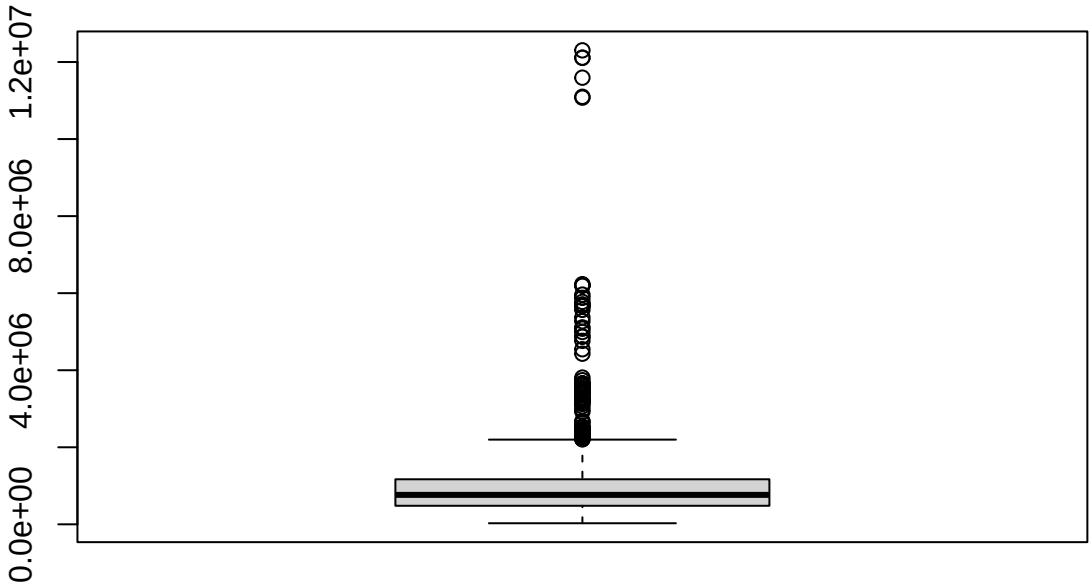

```
boxplot(data$pm10_kg)
```



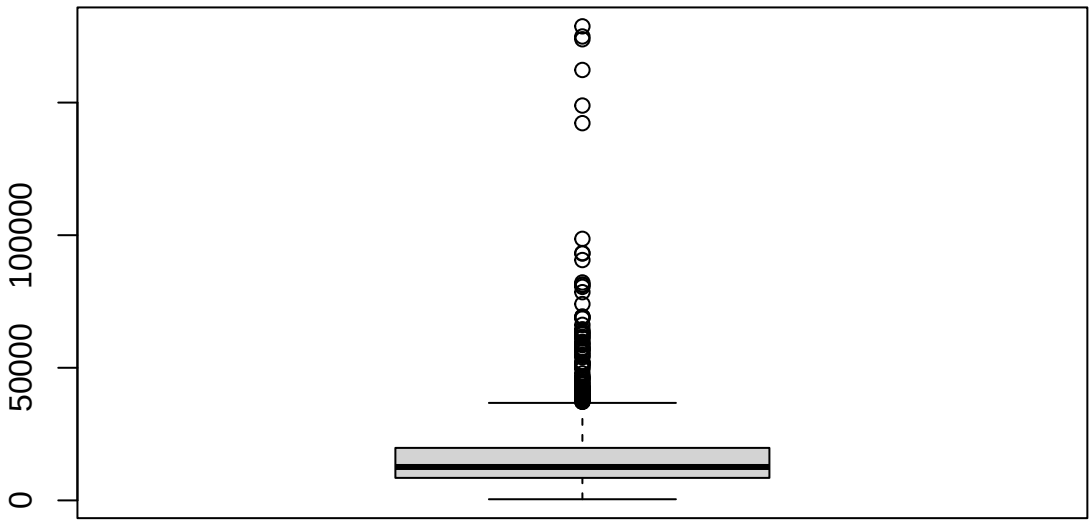

```
boxplot(data$pm25_kg)
```



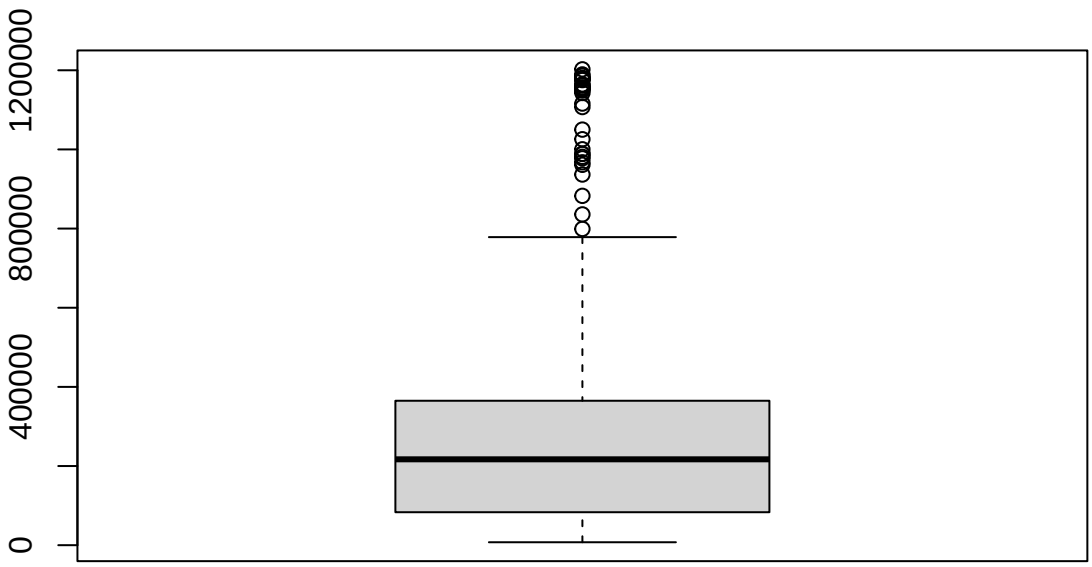

```
boxplot(data$co_kg)
```



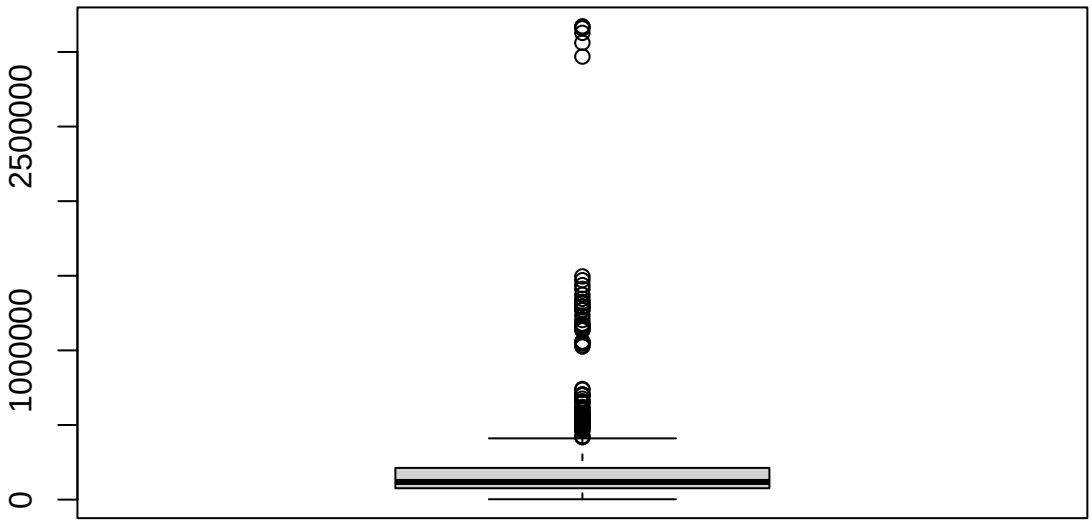

```
boxplot(data$c6h6_kg)
```



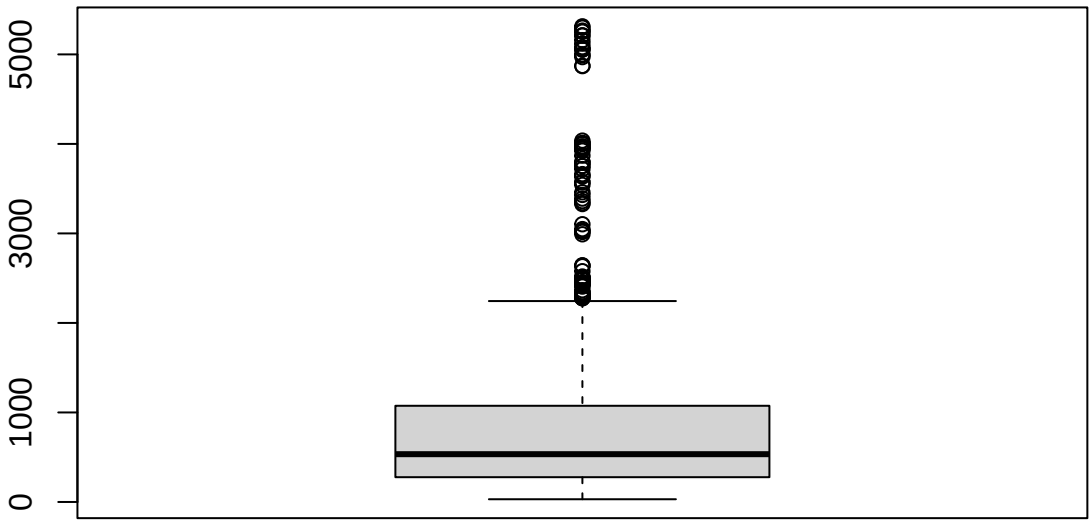

```
boxplot(data$nh3_kg)
```



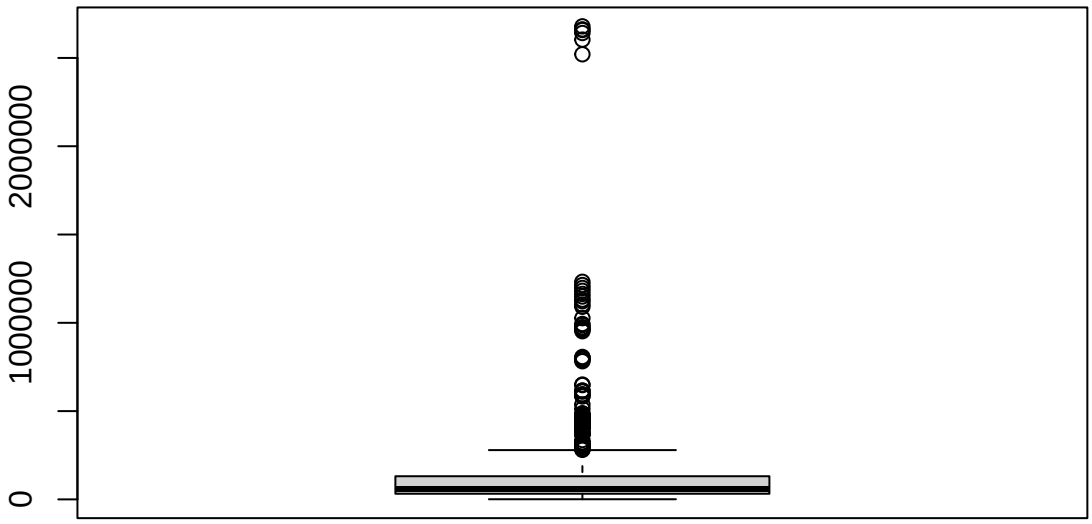

```
boxplot(data$ges_teqco2)
```



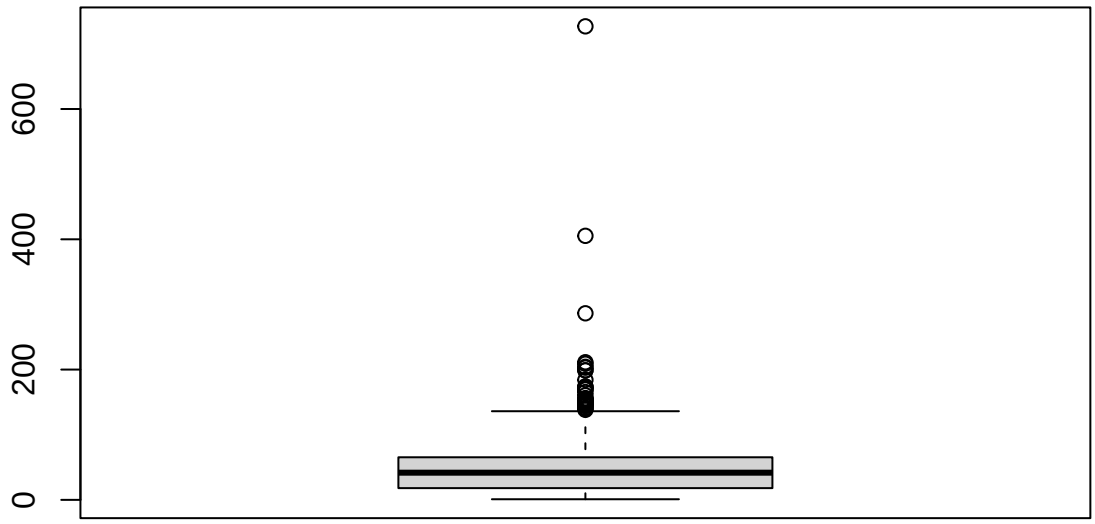

```
boxplot(data$ch4_t)
```



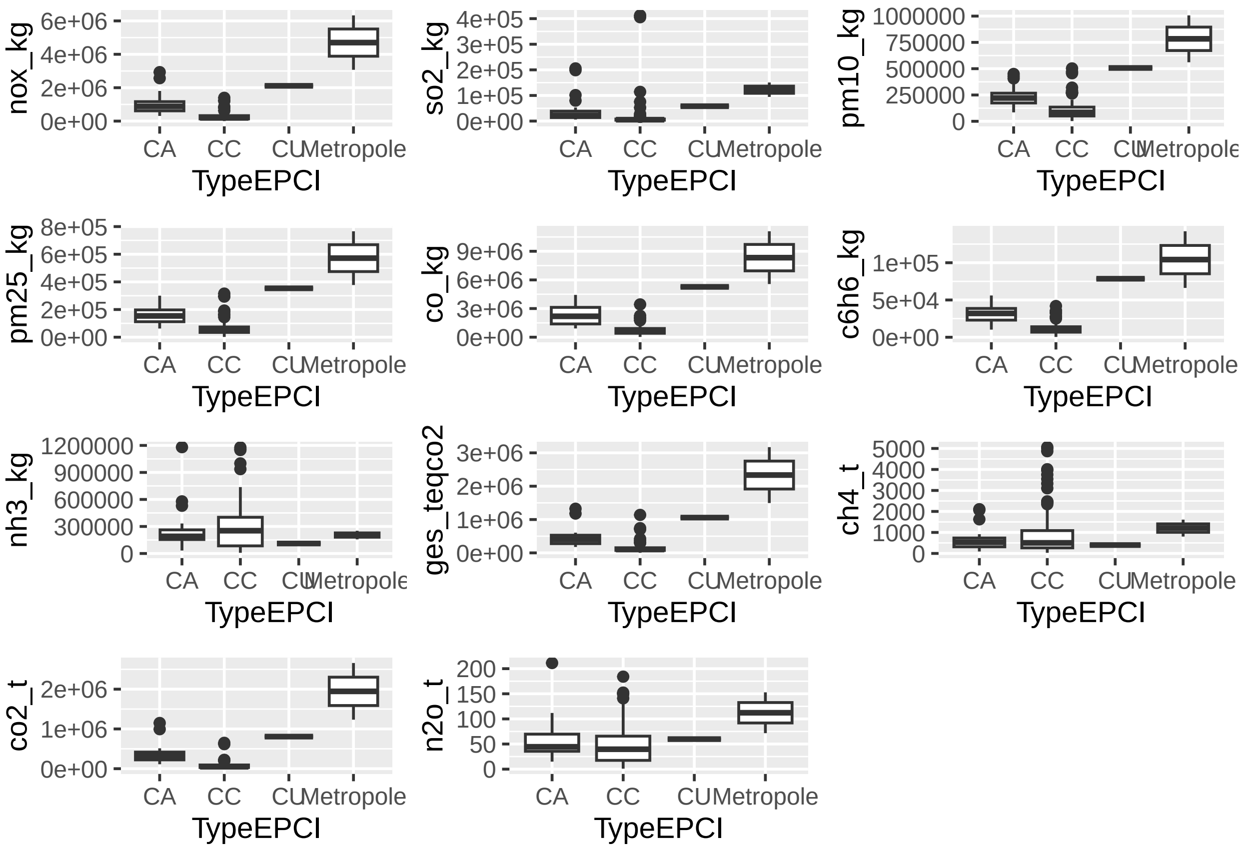

```
boxplot(data$co2_t)
```




```
boxplot(data$n2o_t)
```

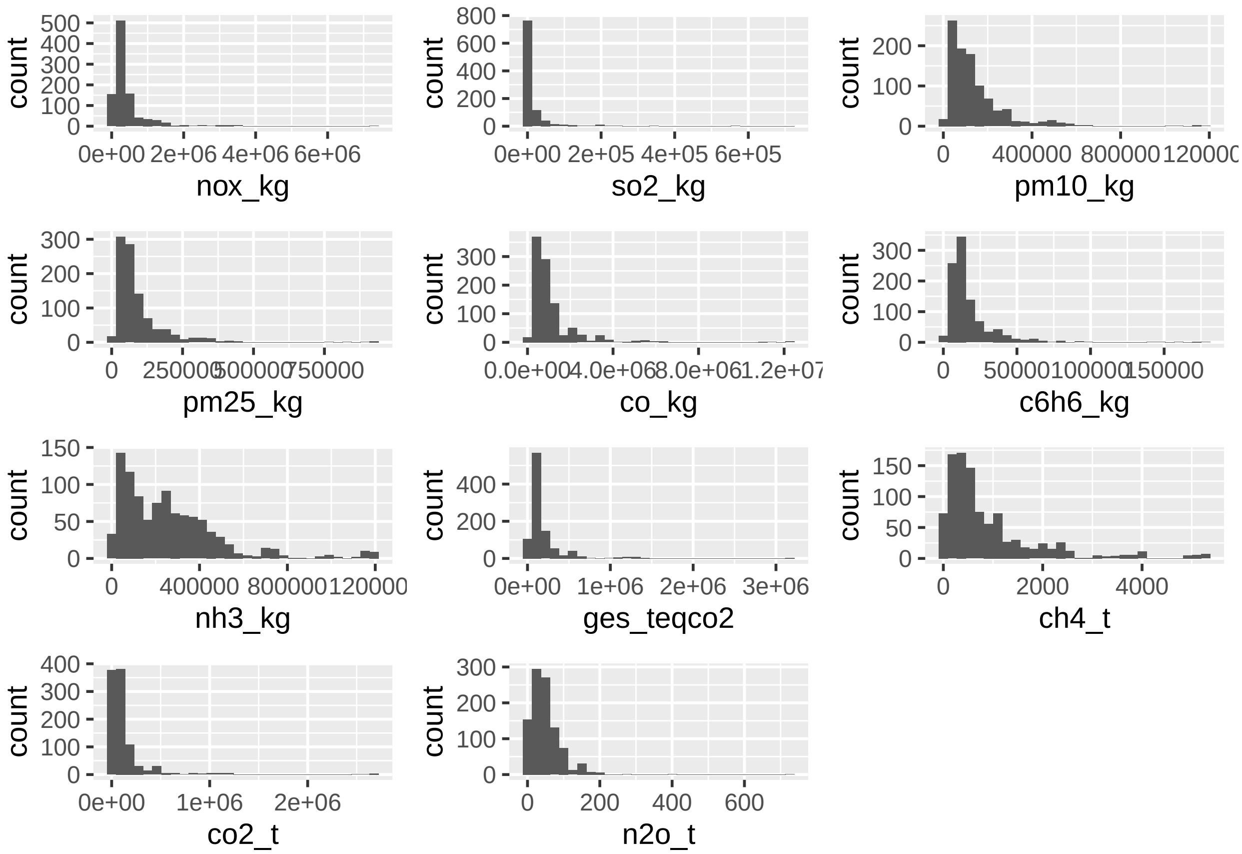



```
data2019= data[1:164,]
g1 = ggplot(data2019, aes(x = TypeEPCI, y = nox_kg)) + geom_boxplot()
g2 = ggplot(data2019, aes(x = TypeEPCI, y = so2_kg)) + geom_boxplot()
g3 = ggplot(data2019, aes(x = TypeEPCI, y = pm10_kg)) + geom_boxplot()
g4 = ggplot(data2019, aes(x = TypeEPCI, y = pm25_kg)) + geom_boxplot()
g5 = ggplot(data2019, aes(x = TypeEPCI, y = co_kg)) + geom_boxplot()
g6 = ggplot(data2019, aes(x = TypeEPCI, y = c6h6_kg)) + geom_boxplot()
g7 = ggplot(data2019, aes(x = TypeEPCI, y = ges_teqco2)) + geom_boxplot()
g8 = ggplot(data2019, aes(x = TypeEPCI, y = ch4_t)) + geom_boxplot()
g9 = ggplot(data2019, aes(x = TypeEPCI, y = co2_t)) + geom_boxplot()
g10 = ggplot(data2019, aes(x = TypeEPCI, y = n2o_t)) + geom_boxplot()
g11 = ggplot(data2019, aes(x = TypeEPCI, y = nh3_kg)) + geom_boxplot()
grid.arrange(g1, g2, g3, g4, g5, g6, g11, g7, g8, g9, g10, ncol = 3)
```



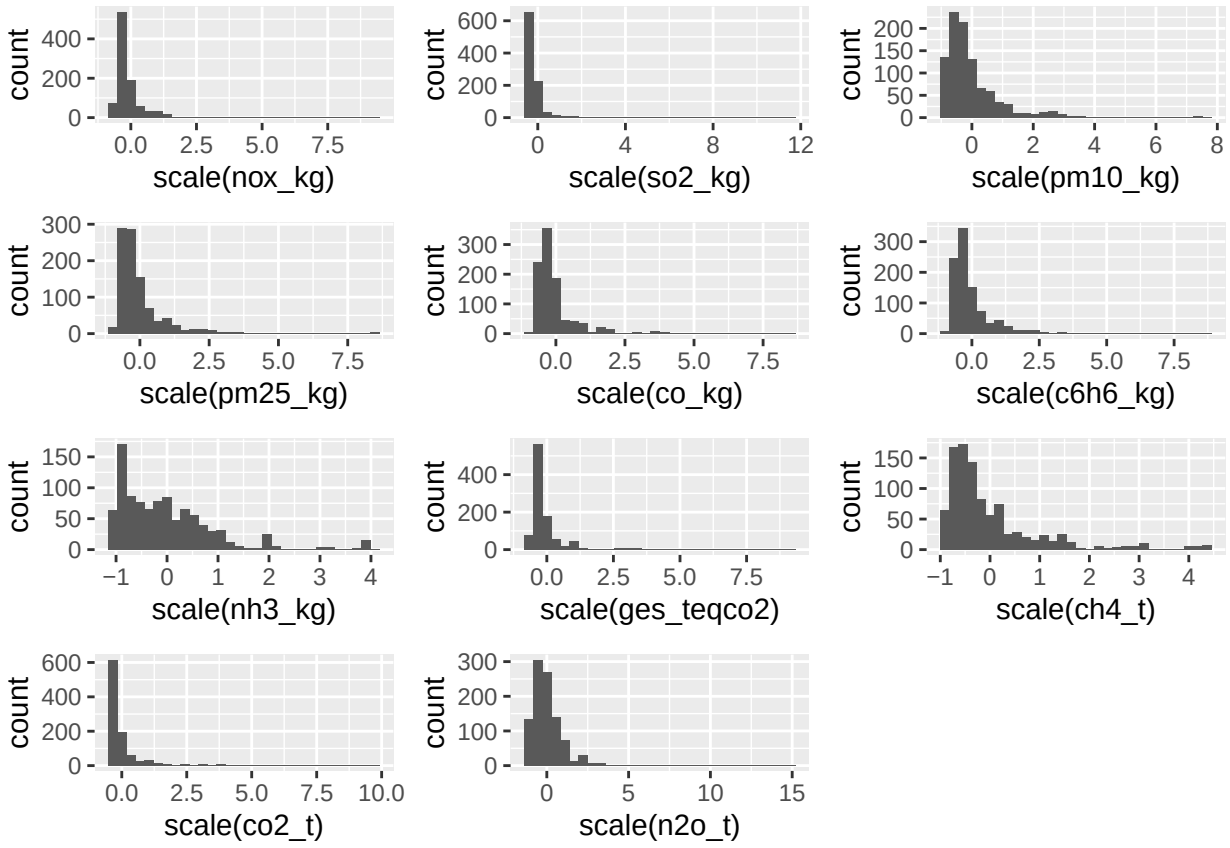

```
g1= ggplot(data = data) + geom_histogram(aes(x = nox_kg))
g2= ggplot(data = data) + geom_histogram(aes(x = so2_kg))
g3= ggplot(data = data) + geom_histogram(aes(x = pm10_kg))
g4= ggplot(data = data) + geom_histogram(aes(x = pm25_kg))
g5= ggplot(data = data) + geom_histogram(aes(x = co_kg))
g6= ggplot(data = data) + geom_histogram(aes(x = c6h6_kg))
g7= ggplot(data = data) + geom_histogram(aes(x = ges_teqco2))
g8= ggplot(data = data) + geom_histogram(aes(x = ch4_t))
g9= ggplot(data = data) + geom_histogram(aes(x = co2_t))
g10= ggplot(data = data) + geom_histogram(aes(x = n2o_t))
g11= ggplot(data = data) + geom_histogram(aes(x = nh3_kg))
grid.arrange(g1, g2, g3, g4, g5, g6, g11, g7, g8, g9, g10, ncol = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
g1= ggplot(data = data) + geom_histogram(aes(x = scale(nox_kg)))
g2= ggplot(data = data) + geom_histogram(aes(x = scale(so2_kg)))
g3= ggplot(data = data) + geom_histogram(aes(x = scale(pm10_kg)))
g4= ggplot(data = data) + geom_histogram(aes(x = scale(pm25_kg)))
g5= ggplot(data = data) + geom_histogram(aes(x = scale(co_kg)))
g6= ggplot(data = data) + geom_histogram(aes(x = scale(c6h6_kg)))
g7= ggplot(data = data) + geom_histogram(aes(x = scale(ges_teqco2)))
g8= ggplot(data = data) + geom_histogram(aes(x = scale(ch4_t)))
g9= ggplot(data = data) + geom_histogram(aes(x = scale(co2_t)))
g10= ggplot(data = data) + geom_histogram(aes(x = scale(n2o_t)))
g11= ggplot(data = data) + geom_histogram(aes(x = scale(nh3_kg)))
grid.arrange(g1, g2, g3, g4, g5, g6, g11, g7, g8, g9, g10, ncol = 3)
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```




```
g1= ggplot(data = data) + geom_histogram(aes(x=log(scale(nox_kg))))
g2= ggplot(data = data) + geom_histogram(aes(x = log(scale(so2_kg))))
g3= ggplot(data = data) + geom_histogram(aes(x = log(scale(pm10_kg))))
g4= ggplot(data = data) + geom_histogram(aes(x = log(scale(pm25_kg))))
g5= ggplot(data = data) + geom_histogram(aes(x = log(scale(co_kg))))
g6= ggplot(data = data) + geom_histogram(aes(x = log(scale(c6h6_kg))))
g7= ggplot(data = data) + geom_histogram(aes(x = log(scale(ges_teqco2))))
g8= ggplot(data = data) + geom_histogram(aes(x = log(scale(ch4_t))))
g9= ggplot(data = data) + geom_histogram(aes(x = log(scale(co2_t))))
g10= ggplot(data = data) + geom_histogram(aes(x = log(scale(n2o_t))))
g11= ggplot(data = data) + geom_histogram(aes(x = log(scale(nh3_kg))))
grid.arrange(g1, g2, g3, g4, g5, g6, g11, g7, g8, g9, g10, ncol = 3)
```

```
## Warning in log(scale(nox_kg)): Production de NaN

## Warning in log(scale(nox_kg)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 741 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(so2_kg)): Production de NaN

## Warning in log(scale(so2_kg)): Production de NaN
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 852 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(pm10_kg)): Production de NaN

## Warning in log(scale(pm10_kg)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 653 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(pm25_kg)): Production de NaN

## Warning in log(scale(pm25_kg)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 701 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(co_kg)): Production de NaN

## Warning in log(scale(co_kg)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 715 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(c6h6_kg)): Production de NaN

## Warning in log(scale(c6h6_kg)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 689 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(nh3_kg)): Production de NaN

## Warning in log(scale(nh3_kg)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 586 rows containing non-finite values (`stat_bin()`).

## Warning in log(scale(ges_teqco2)): Production de NaN

## Warning in log(scale(ges_teqco2)): Production de NaN

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 734 rows containing non-finite values (‘stat_bin()‘).

## Warning in log(scale(ch4_t)): Production de NaN

## Warning in log(scale(ch4_t)): Production de NaN

## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.

## Warning: Removed 659 rows containing non-finite values (‘stat_bin()‘).

## Warning in log(scale(co2_t)): Production de NaN

## Warning in log(scale(co2_t)): Production de NaN

## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.

## Warning: Removed 759 rows containing non-finite values (‘stat_bin()‘).

## Warning in log(scale(n2o_t)): Production de NaN

## Warning in log(scale(n2o_t)): Production de NaN

## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.

## Warning: Removed 577 rows containing non-finite values (‘stat_bin()‘).
```