Ejercicio_cereales

Silvia Pineda

Lectura Fichero de datos

```
datos <- read.csv("CEREALES.csv")  # import data
str(datos)
```

```
'data.frame':   173 obs. of  7 variables:
 $ VARIEDAD : chr  " AVENA" "TRIGO" "CEBADA" "TRIGO" ...
 $ MANGANESO: num  1.31 1.14 0.61 1.05 1.06 1.1 0.6 1.29 1.12 1.11 ...
 $ CALORIAS : num  151 193 126 197 193 ...
 $ FIBRA    : num  4.04 8.01 14.22 7.93 8.01 ...
 $ SELENIO  : num  19.1 24.8 36 24.9 25.8 ...
 $ FOSFORO  : num  178 120 224 120 113 ...
 $ N_MUESTRA: int  1 2 3 4 5 6 7 8 9 10 ...
```

```
summary(datos)
```

```
   VARIEDAD           MANGANESO       CALORIAS         FIBRA       
 Length:173         Min.   :0.58   Min.   :118.1   Min.   : 0.720  
 Class :character   1st Qu.:1.04   1st Qu.:148.1   1st Qu.: 4.180  
 Mode  :character   Median :1.11   Median :150.8   Median : 7.970  
                    Mean   :1.10   Mean   :163.6   Mean   : 7.931  
                    3rd Qu.:1.29   3rd Qu.:195.2   3rd Qu.: 8.190  
                    Max.   :3.84   Max.   :200.4   Max.   :15.210  
    SELENIO         FOSFORO        N_MUESTRA     
 Min.   :18.20   Min.   :105.9   Min.   :  1.0  
 1st Qu.:19.18   1st Qu.:121.9   1st Qu.: 43.0  
 Median :24.77   Median :164.5   Median : 85.0  
 Mean   :25.98   Mean   :162.3   Mean   : 85.3  
 3rd Qu.:26.78   3rd Qu.:177.4   3rd Qu.:128.0  
 Max.   :66.00   Max.   :255.2   Max.   :170.0  
```

```r
datos$VARIEDAD<-factor(datos$VARIEDAD)
str(datos)
```

```
'data.frame':   173 obs. of  7 variables:
 $ VARIEDAD : Factor w/ 3 levels " AVENA","CEBADA",..: 1 3 2 3 3 3 2 1 3 3 ...
 $ MANGANESO: num  1.31 1.14 0.61 1.05 1.06 1.1 0.6 1.29 1.12 1.11 ...
 $ CALORIAS : num  151 193 126 197 193 ...
 $ FIBRA    : num  4.04 8.01 14.22 7.93 8.01 ...
 $ SELENIO  : num  19.1 24.8 36 24.9 25.8 ...
 $ FOSFORO  : num  178 120 224 120 113 ...
 $ N_MUESTRA: int  1 2 3 4 5 6 7 8 9 10 ...
```

```r
table(datos$VARIEDAD)
```

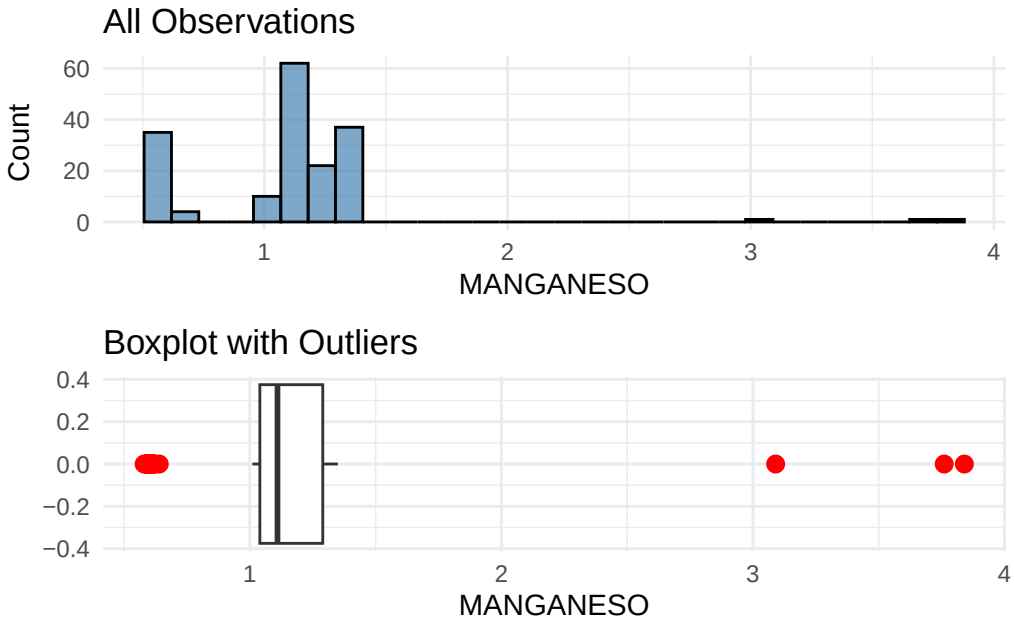
```
 AVENA CEBADA  TRIGO
    61     40     72
```

Uso de la función outliers() y extreme()

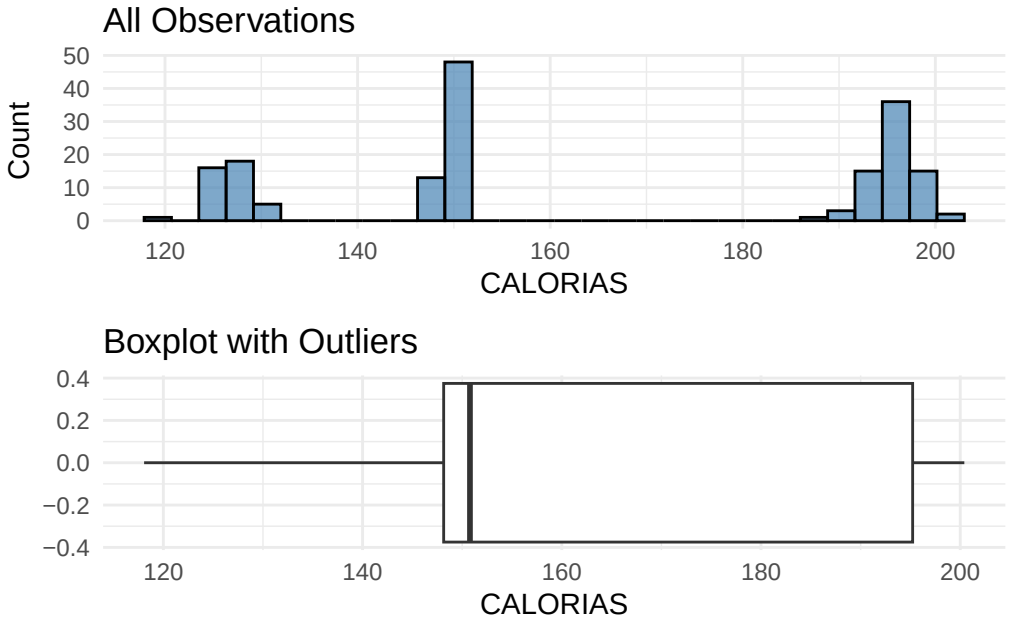
```r
source("outliers.R")
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

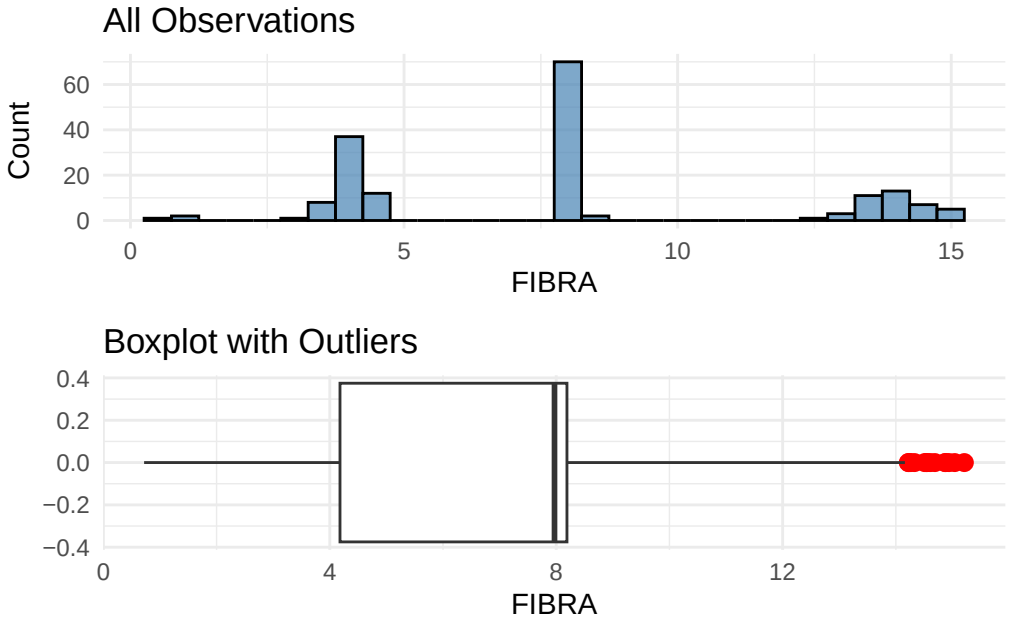
```r
# Aplicar la función a múltiples variables numéricas o enteras
numeric_integer_vars <- names(which(sapply(datos, is.numeric) | sapply(datos, is.integer)))
# Aplicar la función 'outliers' a cada una de las variables numéricas
outliers_results <- lapply(numeric_integer_vars, function(var) {
  outliers(datos, var)  # Llamar a la función pasando el nombre de la variable
})
```



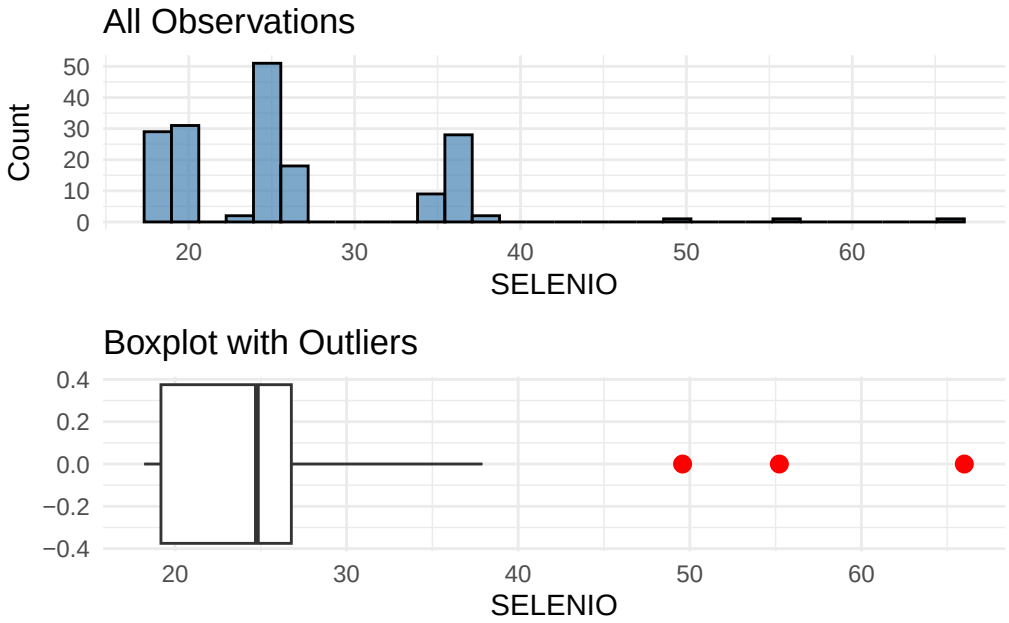

```
Outliers identified in MANGANESO :  42 outliers
Proportion (%) of outliers: 24.28 %
```



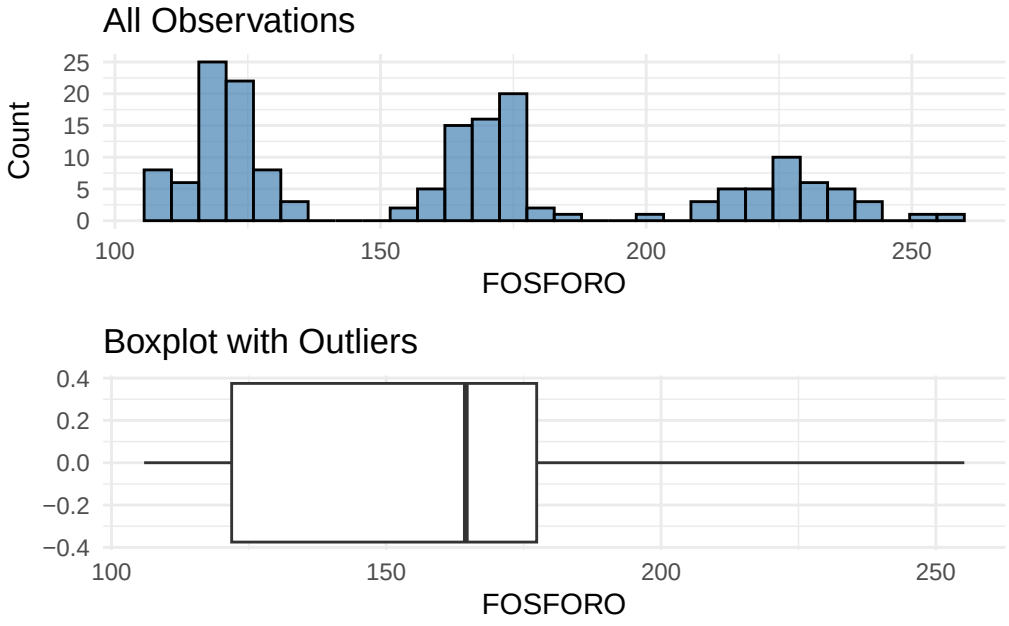
```
Outliers identified in CALORIAS :  0 outliers
Proportion (%) of outliers: 0 %
```



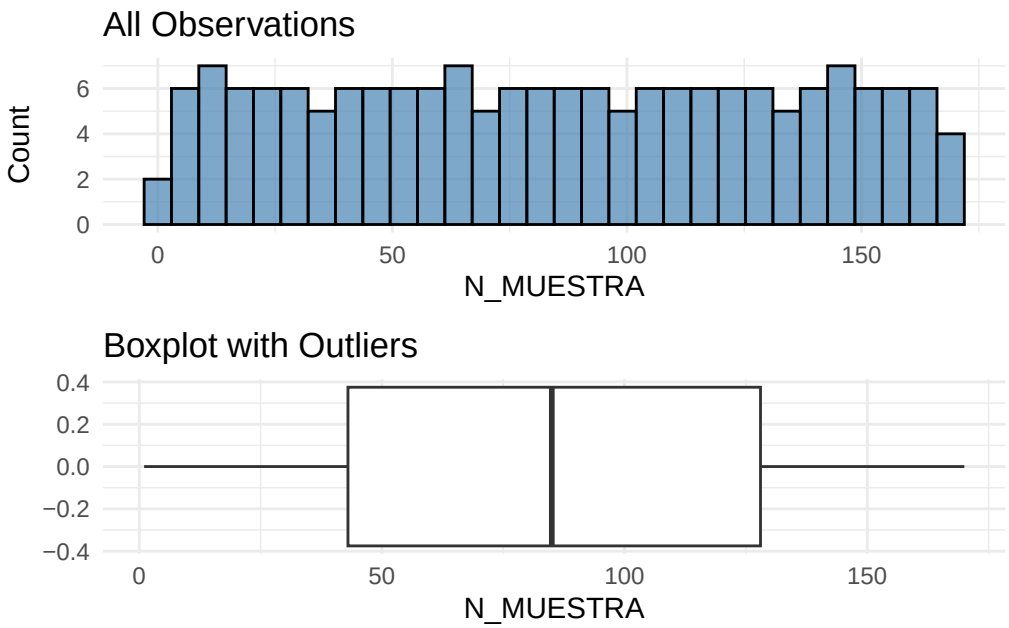

```
Outliers identified in FIBRA :  14 outliers
Proportion (%) of outliers: 8.09 %
```




```
Outliers identified in SELENIO :  3 outliers
Proportion (%) of outliers: 1.73 %
```

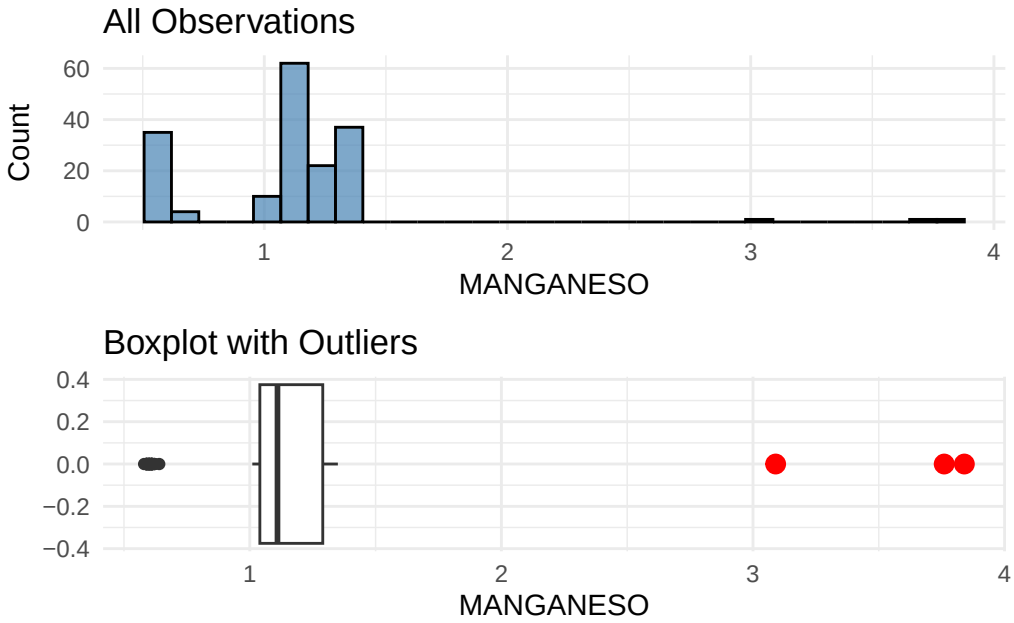


```
Outliers identified in FOSFORO :  0 outliers
Proportion (%) of outliers: 0 %
```

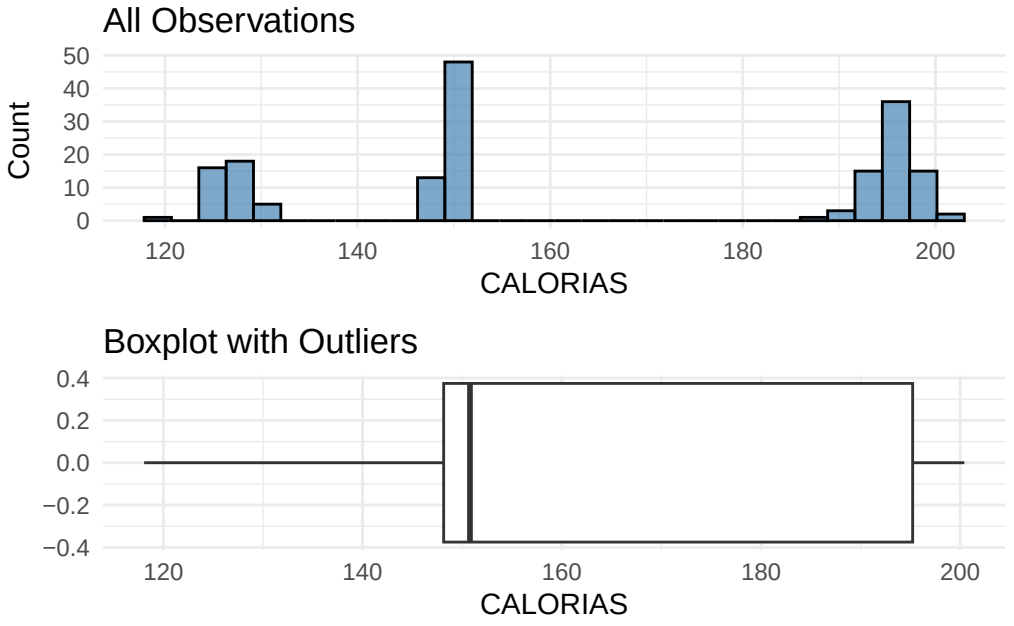



```
Outliers identified in N_MUESTRA :  0 outliers
Proportion (%) of outliers: 0 %
```

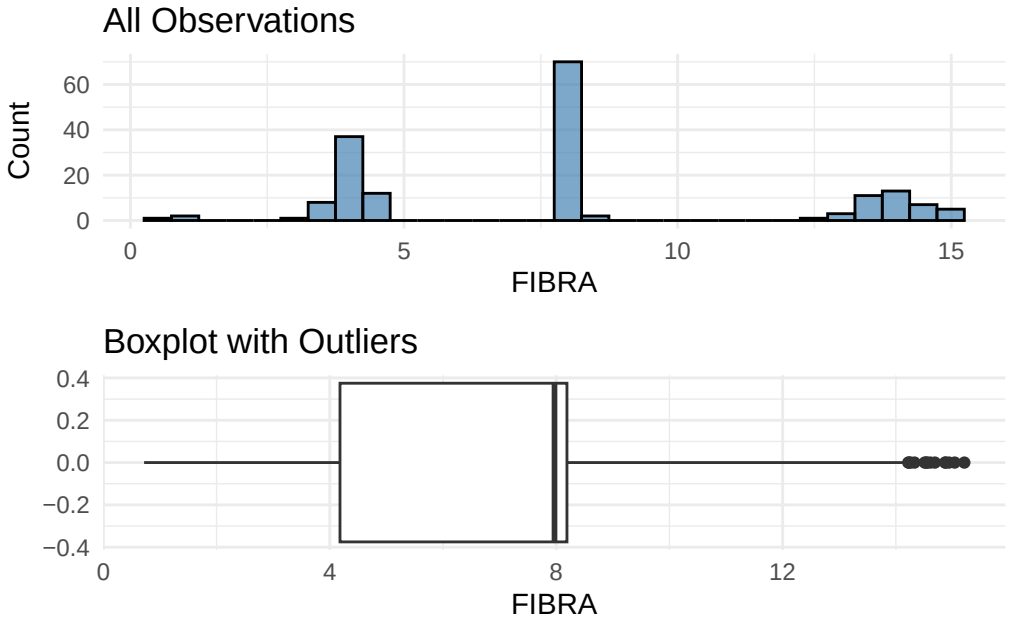
```
extreme_results <- lapply(numeric_integer_vars, function(var) {
  extreme(datos, var)  # Llamar a la función pasando el nombre de la variable
})
```



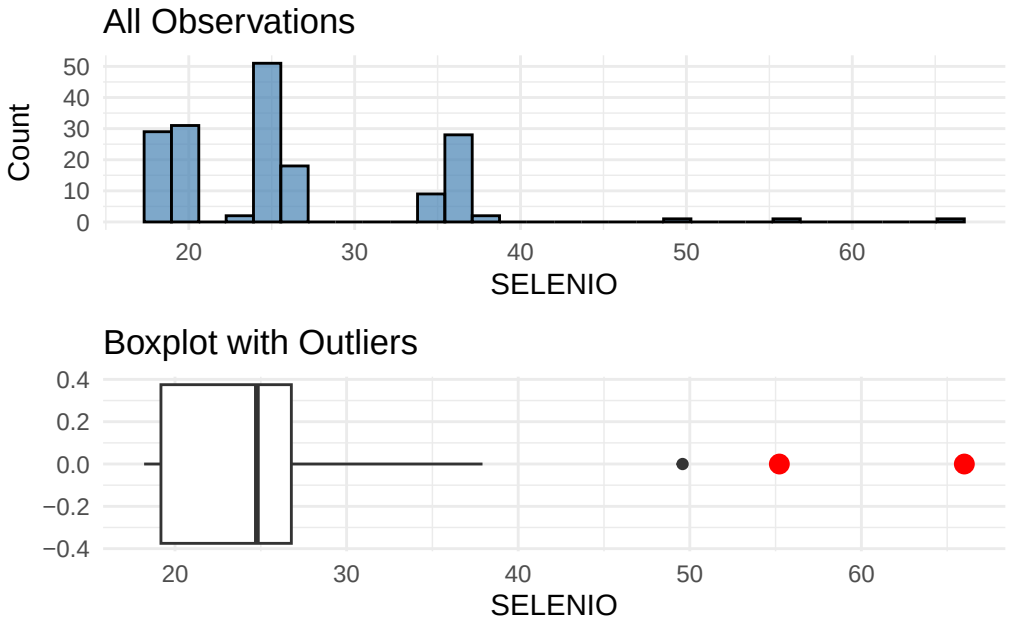

```
Outliers identified in MANGANESO :  3 outliers
Proportion (%) of outliers: 1.73 %
```



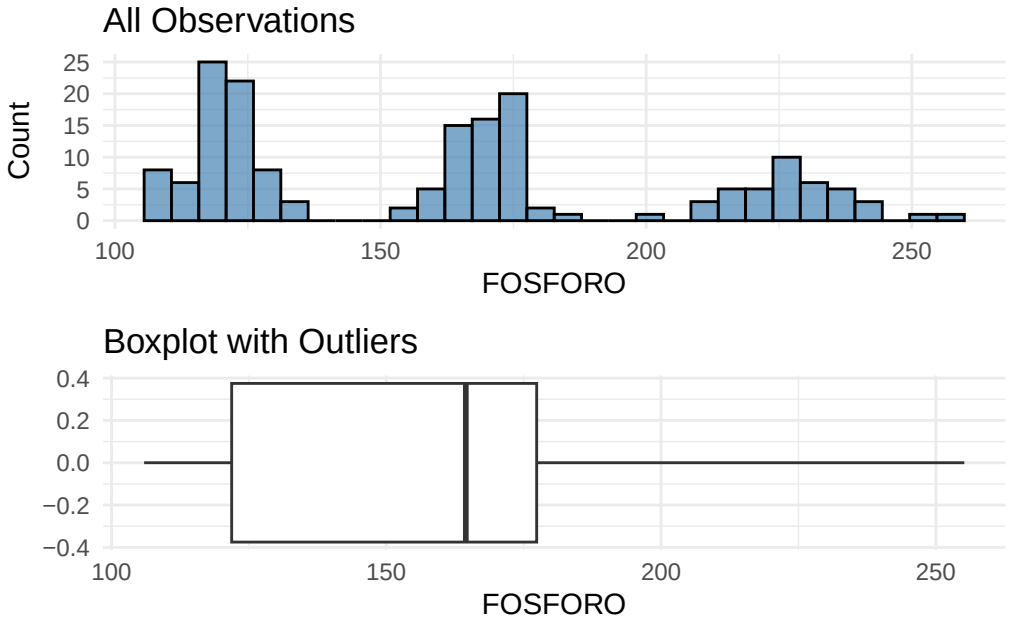
```
Outliers identified in CALORIAS :  0 outliers
Proportion (%) of outliers: 0 %
```



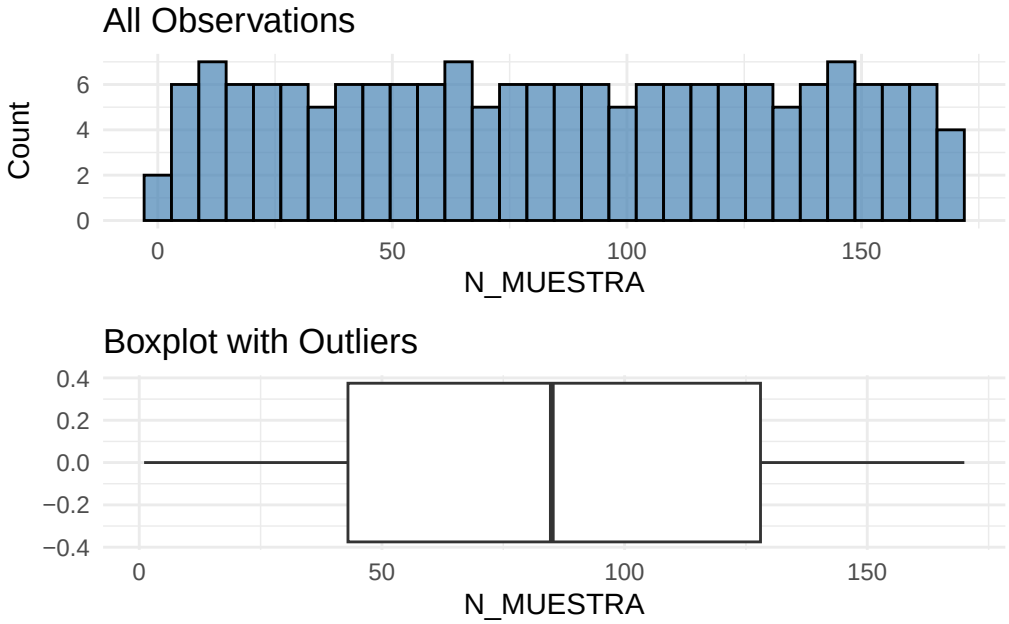

```
Outliers identified in FIBRA :  0 outliers
Proportion (%) of outliers: 0 %
```




```
Outliers identified in SELENIO :  2 outliers
Proportion (%) of outliers: 1.16 %
```



```
Outliers identified in FOSFORO :  0 outliers
Proportion (%) of outliers: 0 %
```




```
Outliers identified in N_MUESTRA :  0 outliers
Proportion (%) of outliers: 0 %
```

Las variables con datos atípicos son:

MANGANESO (24.28%) de outliers y un 1.73% de extremos

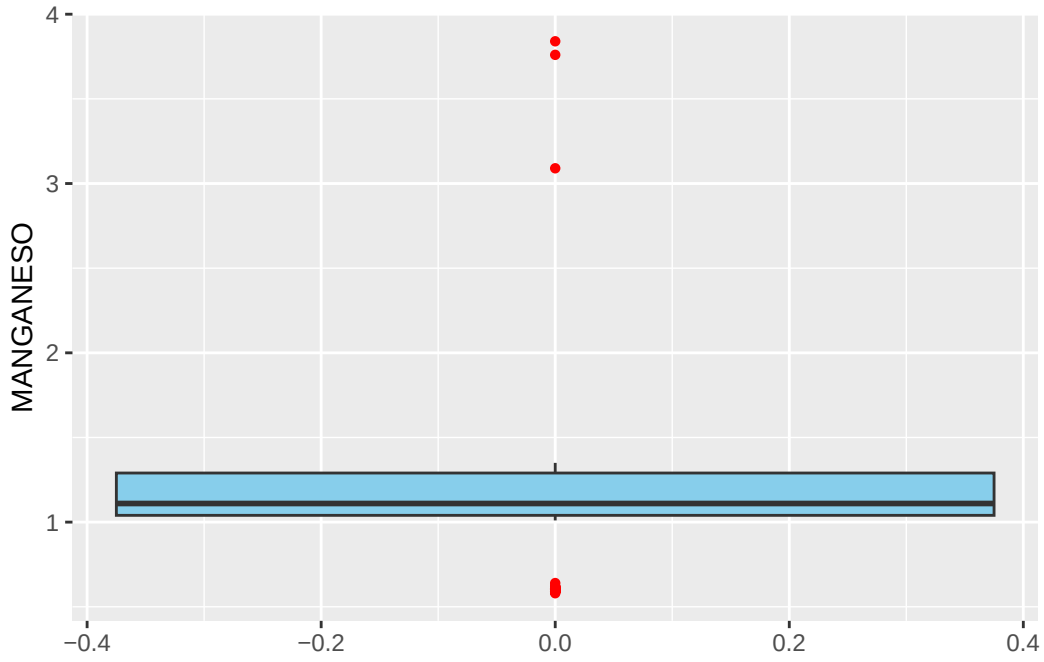
FIBRA (8.09%)

SELENIO (1.73%) de outliers

En todas las composiciones vemos como son distribuciones que claramente se distribuyen en tres grandes grupos que casualmente corresponden a los 3 cereales, así que tendremos que ver que pasa con estas variables y sus posibles outliers / extremos sabiendo que la variable objetivo es VARIEDAD

Estudio de la variable MANGANESO

```
##### MANGANESO #####
ggplot(datos, aes(y = MANGANESO)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16)
```




```
###Los valores atípicos son:
outlier_values <- boxplot.stats(datos$MANGANESO)$out  # outlier values.
out_ind <- which(datos$MANGANESO %in% c(outlier_values))
datos[out_ind,]
```

	VARIEDAD	MANGANESO	CALORIAS	FIBRA	SELENIO	FOSFORO	N_MUESTRA
3	CEBADA	0.61	125.51	14.22	36.02	223.60	3
7	CEBADA	0.60	125.79	14.55	35.62	241.42	7
14	CEBADA	0.61	128.85	15.21	37.24	224.16	14
16	CEBADA	0.61	125.91	14.10	35.93	223.52	16
18	CEBADA	0.62	125.96	13.71	36.01	224.21	18
22	CEBADA	0.61	128.38	13.60	36.12	198.93	22
29	CEBADA	0.59	128.49	14.33	37.08	239.49	29
31	CEBADA	0.60	124.93	13.72	36.64	217.40	31
40	CEBADA	0.59	128.05	14.05	36.03	231.20	39

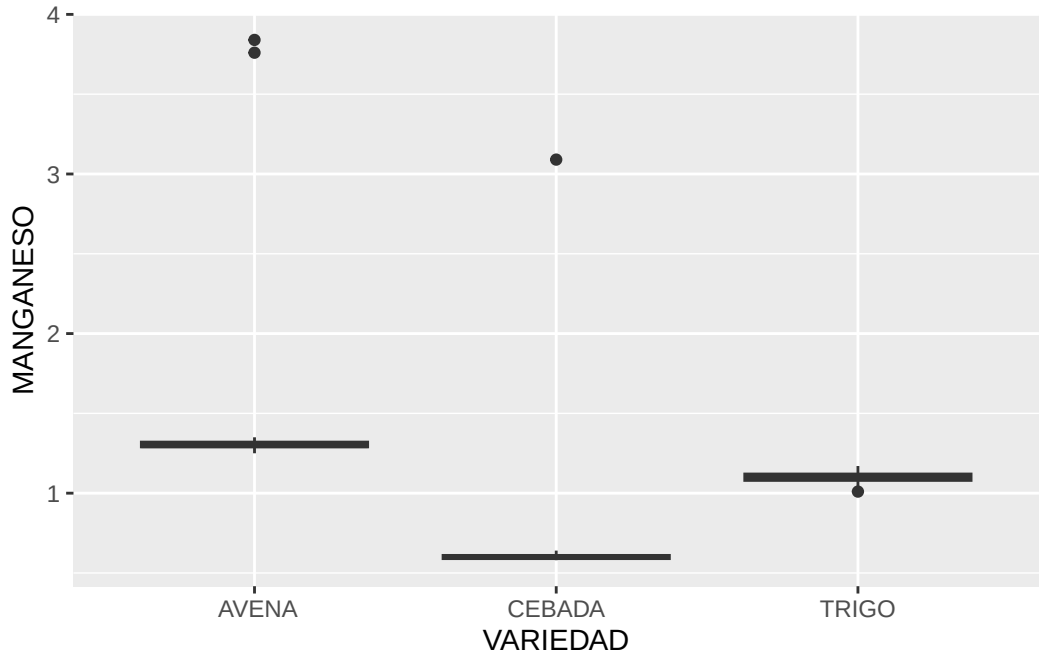
```
41    CEBADA      0.61   129.63 13.76   34.84  223.49        40
48    CEBADA      0.59   129.72 14.51   35.59  211.20        47
50    CEBADA      0.61   126.18 14.88   37.92  227.01        49
52    CEBADA      0.61   127.05 13.38   35.92  223.63        51
54    CEBADA      0.59   118.07 14.69   36.86  227.43        53
61    CEBADA      3.09   130.02 15.04   36.14  232.83        60
68    CEBADA      0.61   128.52 14.26   36.91  234.63        67
69    CEBADA      0.60   130.14 13.19   66.00  228.88        68
77    CEBADA      0.59   125.04 14.61   35.78  217.11        76
85    CEBADA      0.61   129.19 12.91   36.27  224.26        84
87    CEBADA      0.61   124.95 13.44   34.82  209.90        86
88    CEBADA      0.60   126.85 13.75   35.55  212.31        87
97    CEBADA      0.59   124.56 13.76   34.63  237.14        96
98    CEBADA      0.60   127.36 13.41   35.64  227.27        97
99    CEBADA      0.58   124.43 14.87   34.87  250.89        98
105   CEBADA      0.60   124.60 13.63   35.12  224.93       104
108   CEBADA      0.59   126.64 12.63   36.02  232.20       106
110   CEBADA      0.60   126.32 13.99   36.51  226.30       108
112    AVENA      3.76   150.18  0.83   18.99  162.01       110
113   CEBADA      0.59   126.65 13.31   36.03  255.18       111
117   CEBADA      0.60   125.66 14.04   35.40  241.55       115
122   CEBADA      0.61   126.38 14.23   35.90  238.78       120
127    AVENA      3.84   150.51  0.72   18.67  167.37       125
131   CEBADA      0.62   126.83 13.37   34.74  216.66       129
137   CEBADA      0.60   128.03 13.16   35.17  218.92       135
138   CEBADA      0.59   127.53 14.54   36.33  234.38       136
139   CEBADA      0.60   124.55 13.82   36.08  218.57       137
145   CEBADA      0.61   126.32 13.74   35.38  217.58       143
154   CEBADA      0.64   128.94 14.94   35.86  231.59       152
155   CEBADA      0.60   127.24 13.57   36.53  229.52       153
161   CEBADA      0.59   128.60 14.16   36.86  238.80       159
170   CEBADA      0.60   125.52 14.02   36.46  231.55       167
172   CEBADA      0.62   127.74 13.95   35.94  226.46       169
```

```
###Los valores extremos son:
extreme_values <- boxplot.stats(datos$MANGANESO,coef=3)$out  # extreme values.
ext_ind <- which(datos$MANGANESO %in% c(extreme_values))
datos[ext_ind,]
```

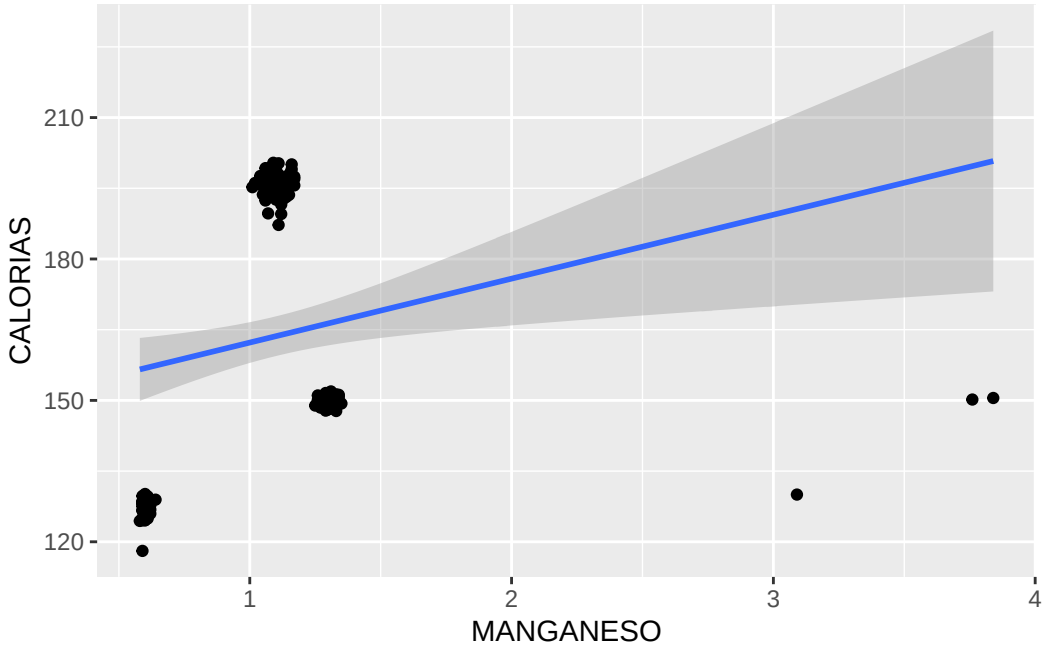
```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA
61    CEBADA      3.09   130.02 15.04   36.14  232.83        60
112    AVENA      3.76   150.18  0.83   18.99  162.01       110
```

```
127    AVENA      3.84   150.51  0.72   18.67  167.37       125
```

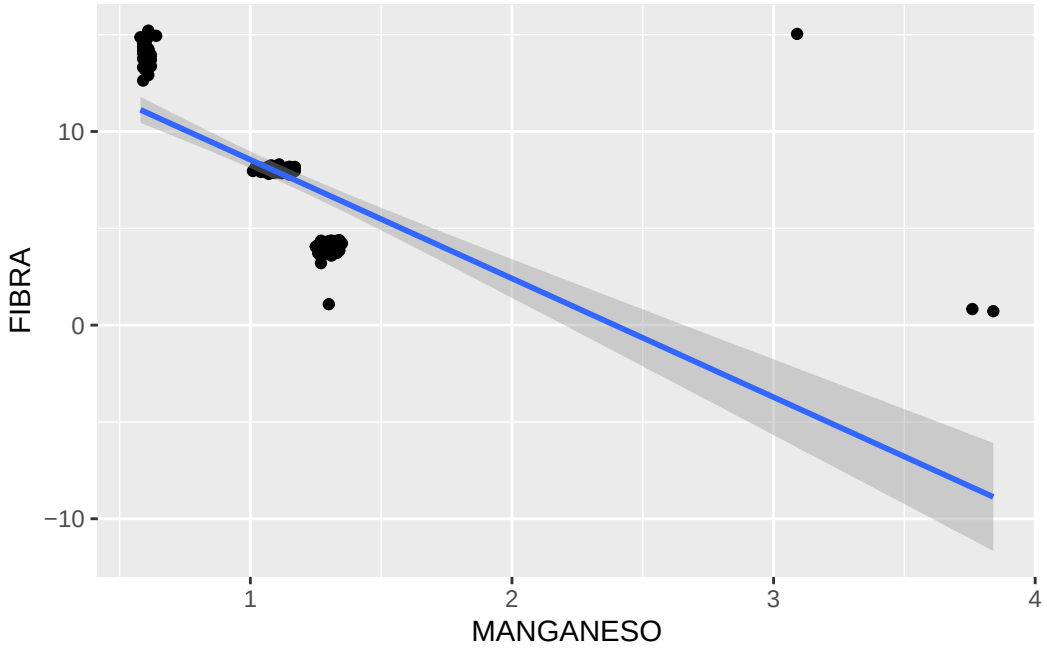
```r
library(patchwork)  # Para combinar gráficos fácilmente

# Gráfico 1: Manganeso por variedad
p1 <- ggplot(datos, aes(x = VARIEDAD, y = MANGANESO)) +
  geom_boxplot(fill = "lightblue")

p1
```



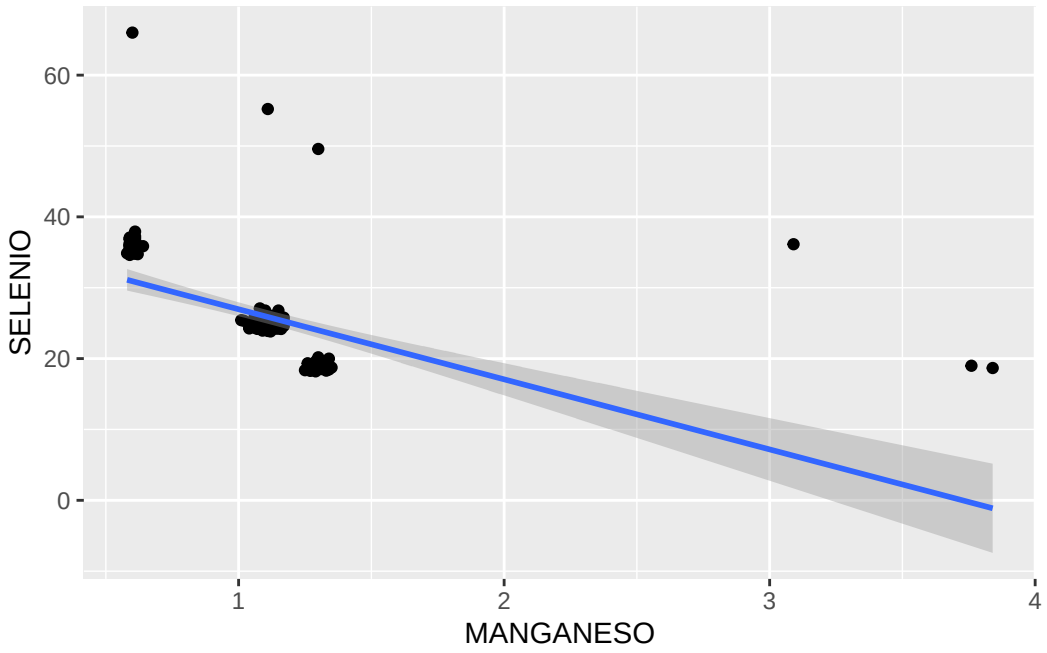

```r
ggp <- ggplot(datos,aes(MANGANESO, CALORIAS)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```



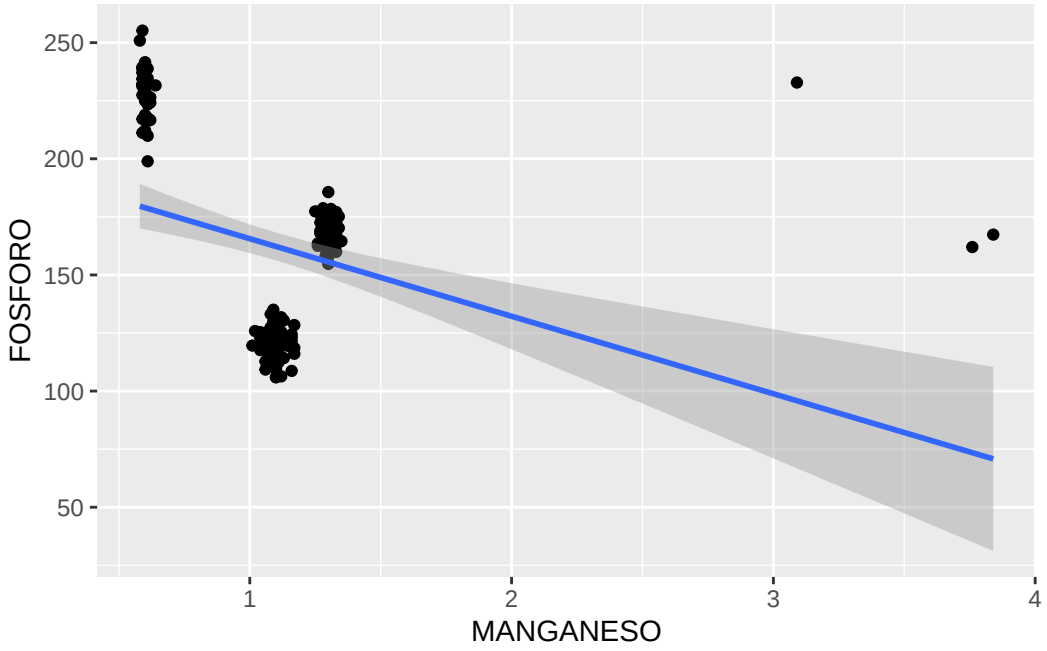

```
ggp <- ggplot(datos,aes(MANGANESO, FIBRA)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```




```
ggp <- ggplot(datos,aes(MANGANESO, SELENIO)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```



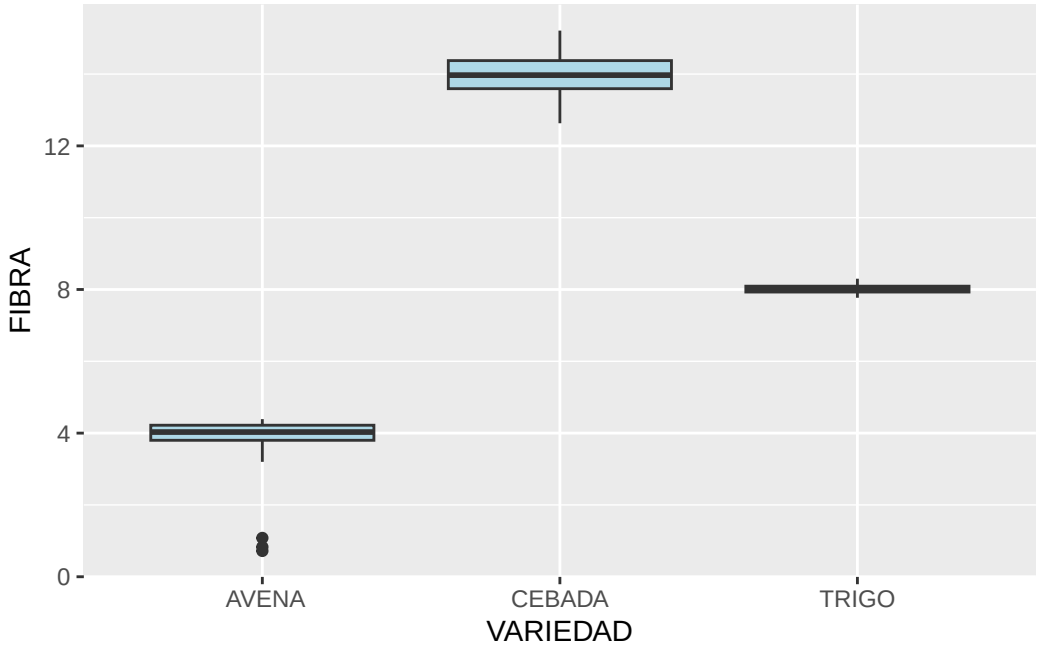

```
ggp <- ggplot(datos,aes(MANGANESO, FOSFORO)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```



Los outliers son un porcentaje muy elevado, pero los extremos corresponden a un 1.73%, y además en la inspección gráfico vemos claramente que son outliers ya que no están asociados con ninguna variable y especialmente con la variable objetivo. Los borraremos.

Estudio de la variable FIBRA

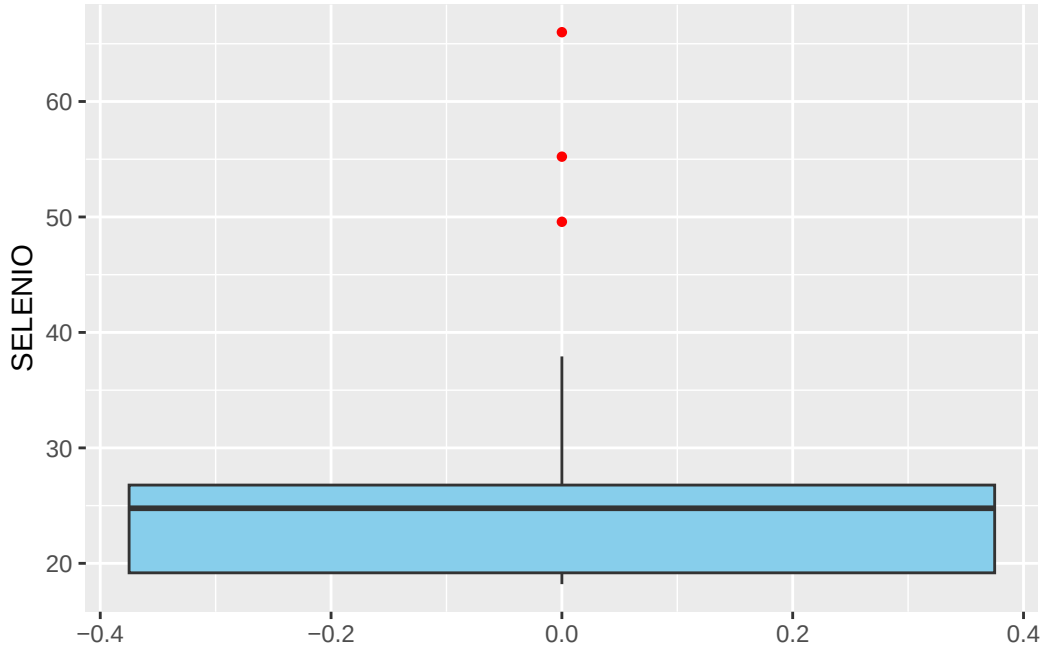
```
# Gráfico 1: Manganeso por variedad
p1 <- ggplot(datos, aes(x = VARIEDAD, y = FIBRA)) +
  geom_boxplot(fill = "lightblue")

p1
```



El porcentaje de ouliers es de un 8% y además se ve claramente que es debido a una distribución asimétrica ya que los outliers corresponden a la cebada, por tanto, no vamos a borrarlos, no son outliers

Estudio de la variable SELENIO

```
##### SELENIO #####
ggplot(datos, aes(y = SELENIO)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16)
```




```
###Los valores atípicos son:
outlier_values <- boxplot.stats(datos$SELENIO)$out  # outlier values.
out_ind <- which(datos$SELENIO %in% c(outlier_values))
datos[out_ind,]
```

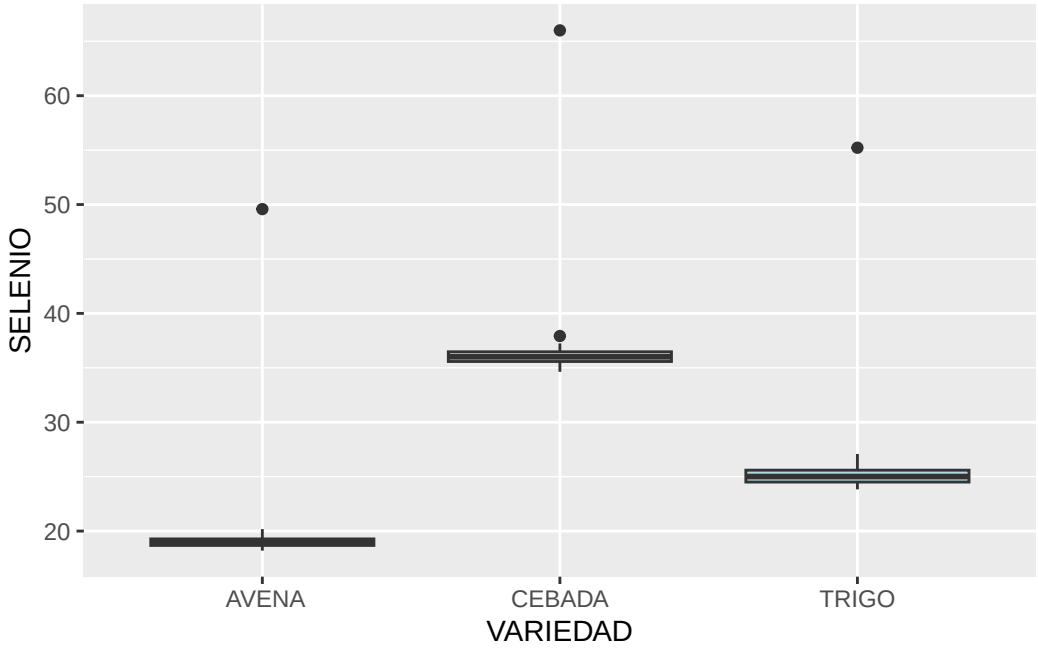
```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA
23     TRIGO      1.11   187.22  7.95   55.22  126.64        23
69    CEBADA      0.60   130.14 13.19   66.00  228.88        68
142    AVENA      1.30   150.18  1.08   49.58  168.59       140
```

```
###Los valores extremos son:
extreme_values <- boxplot.stats(datos$SELENIO,coef=3)$out  # extreme values.
ext_ind <- which(datos$SELENIO %in% c(extreme_values))
datos[ext_ind,]
```

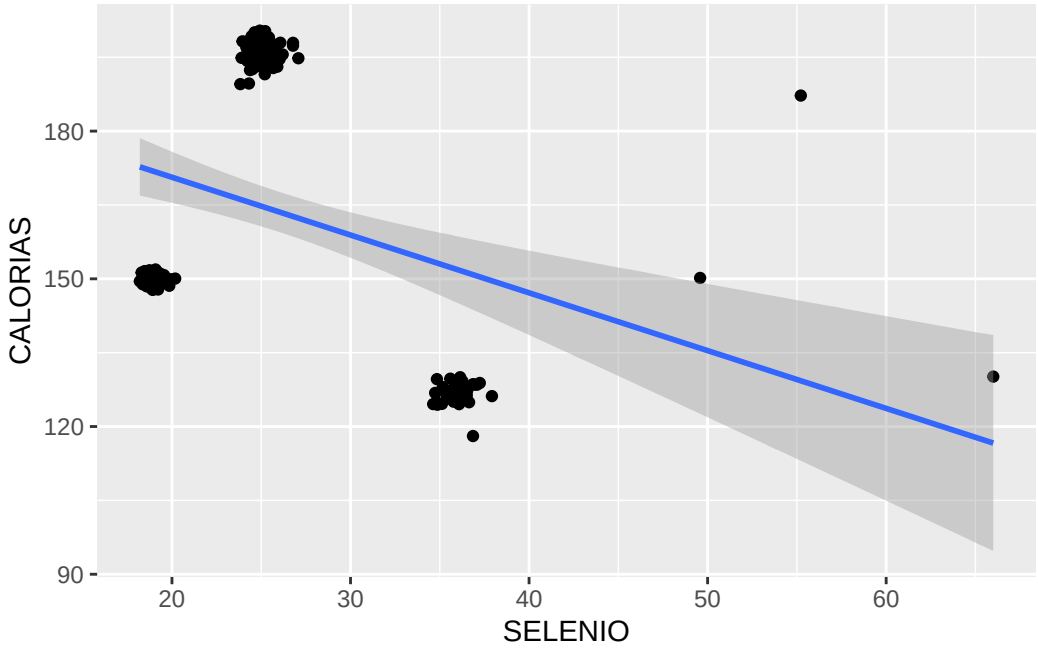
```
   VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA
23    TRIGO      1.11   187.22  7.95   55.22  126.64        23
69   CEBADA      0.60   130.14 13.19   66.00  228.88        68
```

```
# Gráfico 1: Pressure Height por mes
p1 <- ggplot(datos, aes(x = VARIEDAD, y = SELENIO)) +
```

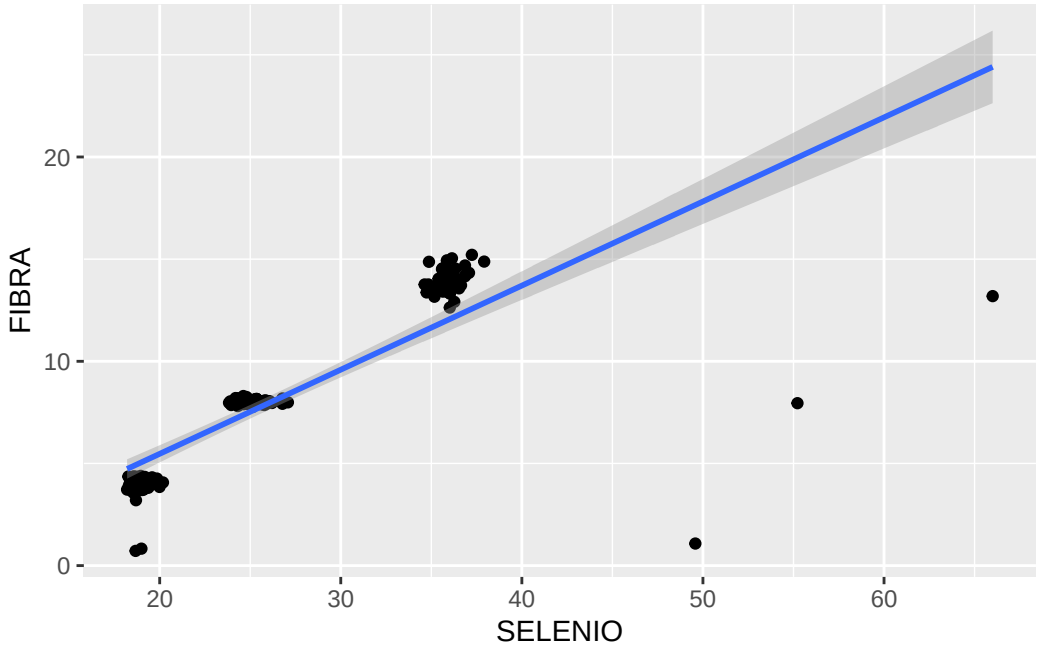
```
  geom_boxplot(fill = "lightblue")

p1
```



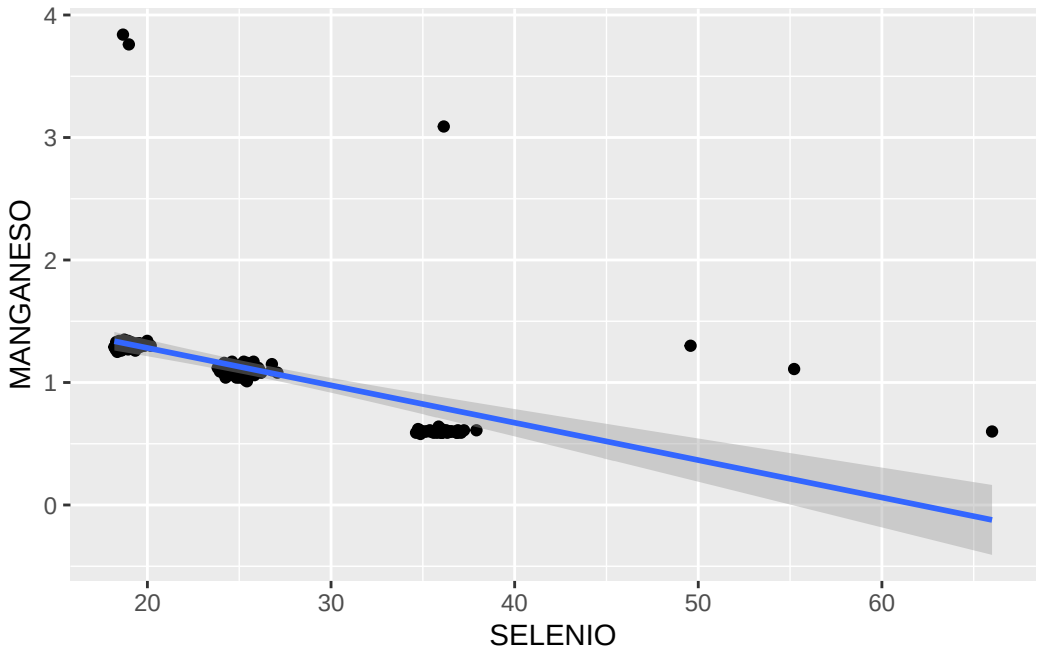

```
ggp <- ggplot(datos,aes(SELENIO, CALORIAS)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```



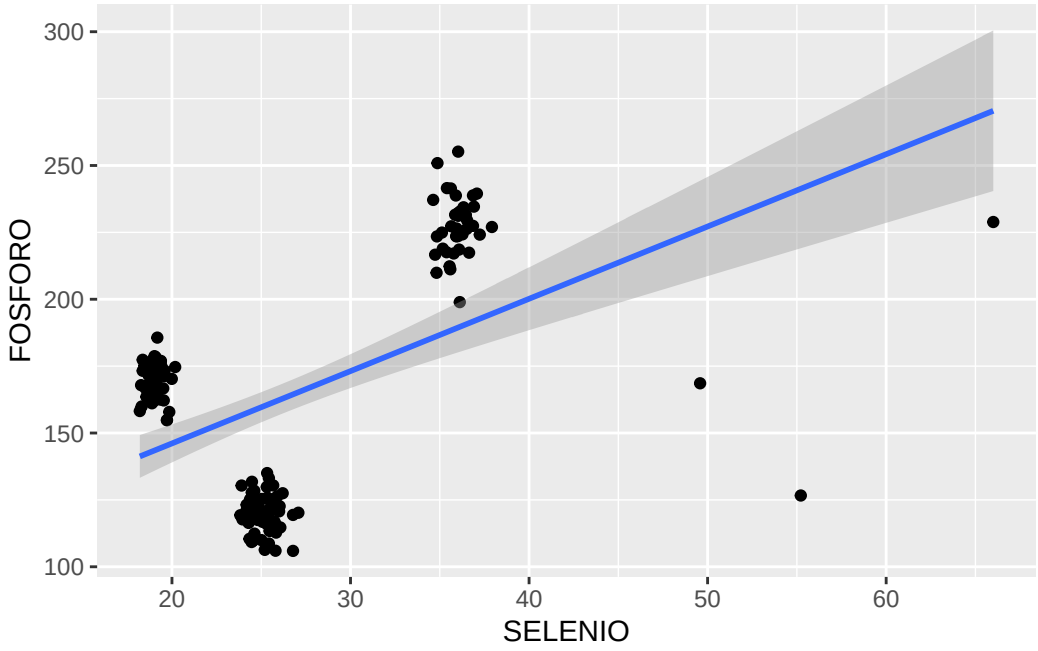

```
ggp <- ggplot(datos,aes(SELENIO, FIBRA)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```




```
ggp <- ggplot(datos,aes(SELENIO, MANGANESO)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```




```
ggp <- ggplot(datos,aes(SELENIO, FOSFORO)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```



En esta variable vemos como los 3 outliers no estáb asociados con la variable objetivo y se salen por completo de todas las nubes de puntos. Los borraremos

CONCLUSIÓN:

Vamos a borrar los outliers de SELENIO y los extremos de manganeso

```
extreme_values <- boxplot.stats(datos$MANGANESO,coef=3)$out  # extreme values.
ext_ind <- which(datos$MANGANESO %in% c(extreme_values))
datos[ext_ind,]
```

```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA
61    CEBADA      3.09   130.02 15.04   36.14  232.83        60
112    AVENA      3.76   150.18  0.83   18.99  162.01       110
127    AVENA      3.84   150.51  0.72   18.67  167.37       125
```

```
datos$MANGANESO[ext_ind]<-NA


outlier_values <- boxplot.stats(datos$SELENIO)$out  # outlier values.
out_ind <- which(datos$SELENIO %in% c(outlier_values))
datos[out_ind,]
```

```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA
23     TRIGO      1.11   187.22  7.95   55.22  126.64        23
69    CEBADA      0.60   130.14 13.19   66.00  228.88        68
142    AVENA      1.30   150.18  1.08   49.58  168.59       140
```

```
datos$SELENIO[out_ind]<-NA

summary(datos)
```

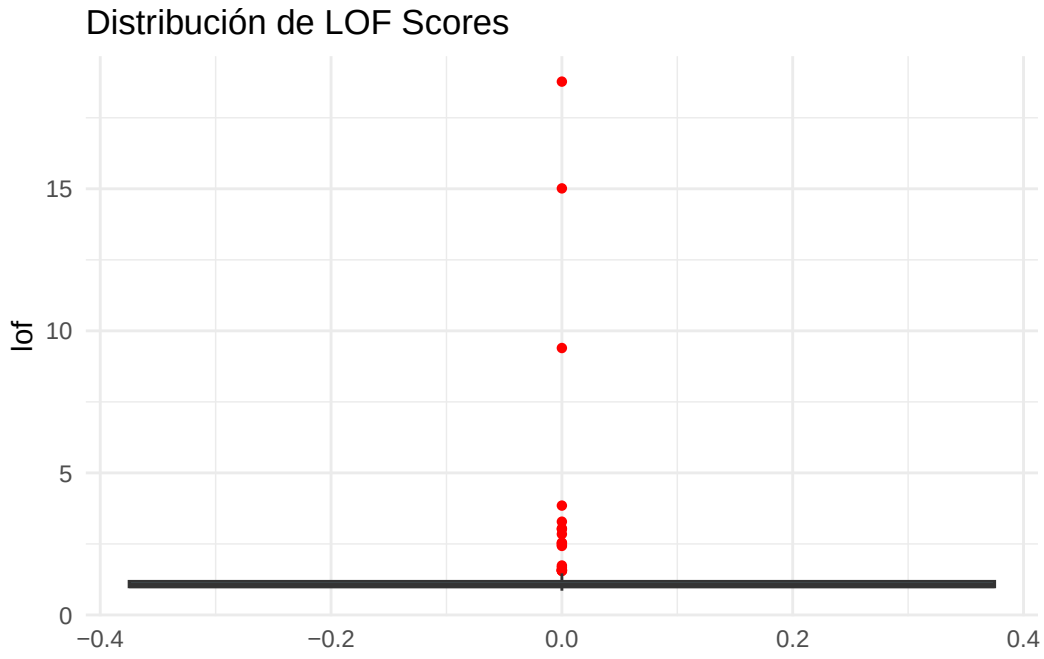
```
   VARIEDAD    MANGANESO        CALORIAS         FIBRA           SELENIO
  AVENA:61   Min.   :0.580   Min.   :118.1   Min.   : 0.720   Min.   :18.20
 CEBADA:40   1st Qu.:1.040   1st Qu.:148.1   1st Qu.: 4.180   1st Qu.:19.17
 TRIGO :72   Median :1.110   Median :150.8   Median : 7.970   Median :24.77
             Mean   :1.056   Mean   :163.6   Mean   : 7.931   Mean   :25.43
             3rd Qu.:1.290   3rd Qu.:195.2   3rd Qu.: 8.190   3rd Qu.:26.17
             Max.   :1.350   Max.   :200.4   Max.   :15.210   Max.   :37.92
             NA's   :3                                        NA's   :3
    FOSFORO        N_MUESTRA
 Min.   :105.9   Min.   :  1.0
 1st Qu.:121.9   1st Qu.: 43.0
 Median :164.5   Median : 85.0
 Mean   :162.3   Mean   : 85.3
 3rd Qu.:177.4   3rd Qu.:128.0
 Max.   :255.2   Max.   :170.0
```

Estudio Multivariante

```
library(dbscan)
```

```
Attaching package: 'dbscan'

The following object is masked from 'package:stats':

    as.dendrogram
```

```r
library(class)
library(ggplot2)

datos <- read.csv("CEREALES.csv")  # import data
datos$VARIEDAD<-factor(datos$VARIEDAD)

####Aplicamos LOF
k<-round(log(nrow(datos)))
lof<-lof(select(datos,-VARIEDAD,-N_MUESTRA),minPts = k)

datos$lof<-lof

ggplot(datos, aes(y = lof)) +
  geom_boxplot(fill = "skyblue", outlier.color = "red", outlier.shape = 16) +
  theme_minimal() +
  labs(title = "Distribución de LOF Scores")
```

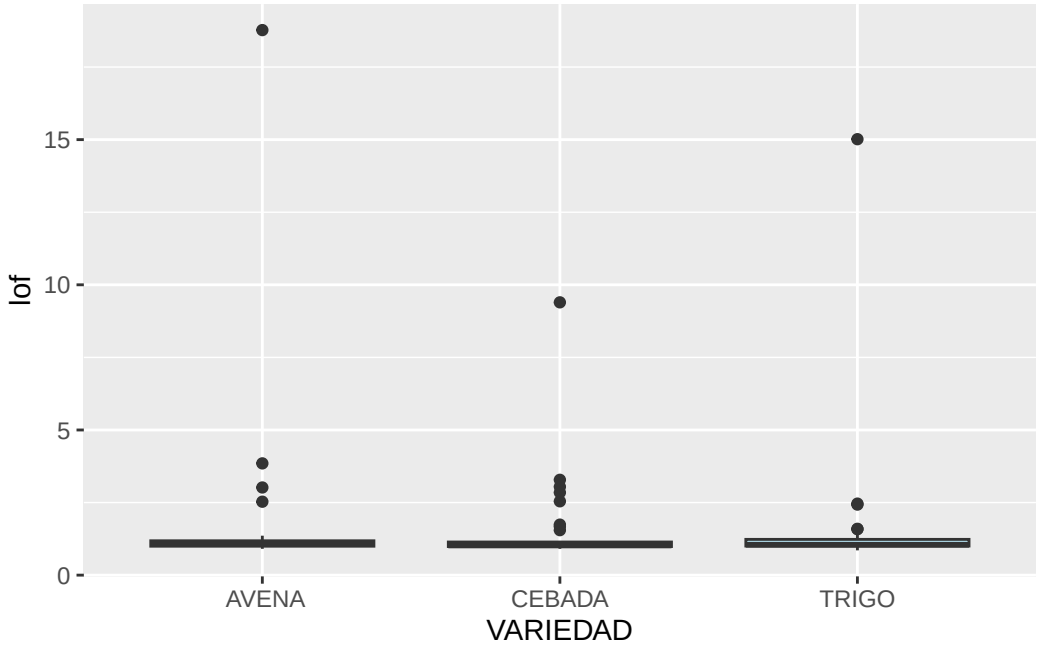



```r
datos[lof>5,]
```

```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA       lof
23     TRIGO      1.11   187.22  7.95   55.22  126.64        23 15.015167
69    CEBADA      0.60   130.14 13.19   66.00  228.88        68  9.395902
```

```
142    AVENA      1.30    150.18  1.08    49.58  168.59        140 18.772921
```

```
####Comprobamos las cualitativas
ggplot(datos, aes(x = VARIEDAD, y = lof)) +
  geom_boxplot(fill = "lightblue")
```




```
###Vamos a ver si son los mismos que los datos que hemos quitado
extreme_values <- boxplot.stats(datos$MANGANESO,coef=3)$out  # extreme values.
ext_ind <- which(datos$MANGANESO %in% c(extreme_values))
datos[ext_ind,]
```

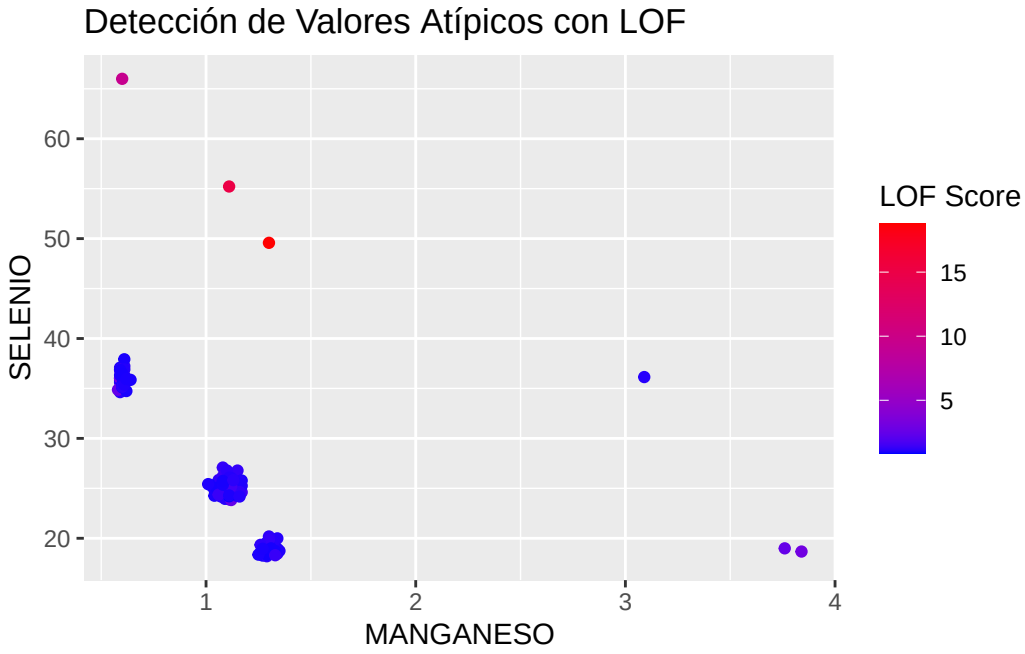
```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA      lof
61    CEBADA      3.09   130.02 15.04   36.14  232.83        60 1.093118
112    AVENA      3.76   150.18  0.83   18.99  162.01       110 2.529654
127    AVENA      3.84   150.51  0.72   18.67  167.37       125 3.019403
```

```
outlier_values <- boxplot.stats(datos$SELENIO)$out  # outlier values.
out_ind <- which(datos$SELENIO %in% c(outlier_values))
datos[out_ind,]
```

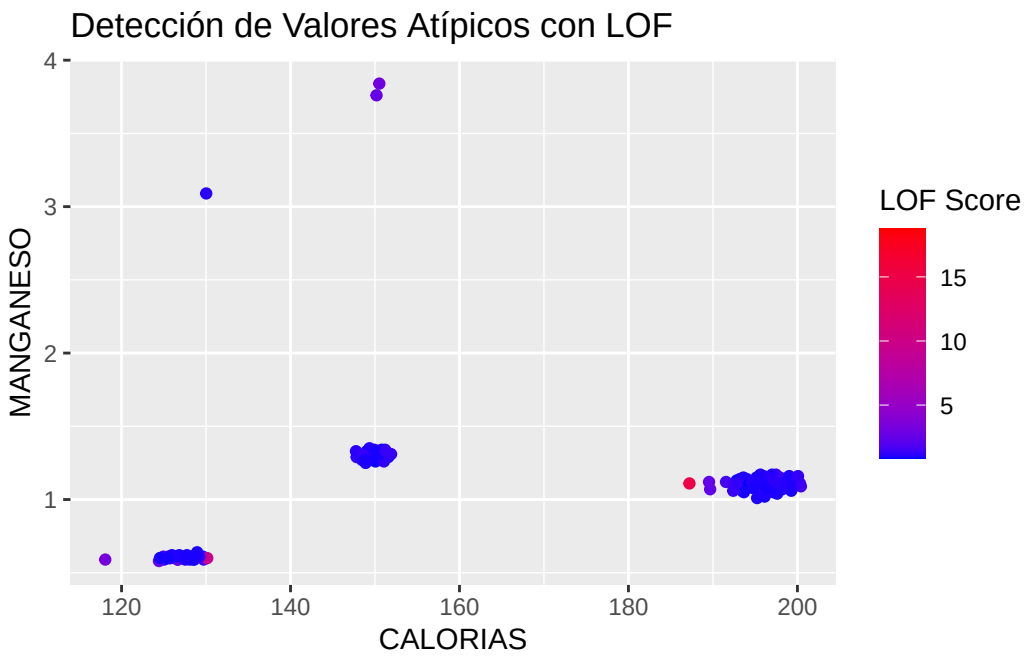
```
    VARIEDAD MANGANESO CALORIAS FIBRA SELENIO FOSFORO N_MUESTRA       lof
```

```
23     TRIGO      1.11   187.22  7.95   55.22  126.64        23 15.015167
69    CEBADA      0.60   130.14 13.19   66.00  228.88        68  9.395902
142    AVENA      1.30   150.18  1.08   49.58  168.59       140 18.772921
```

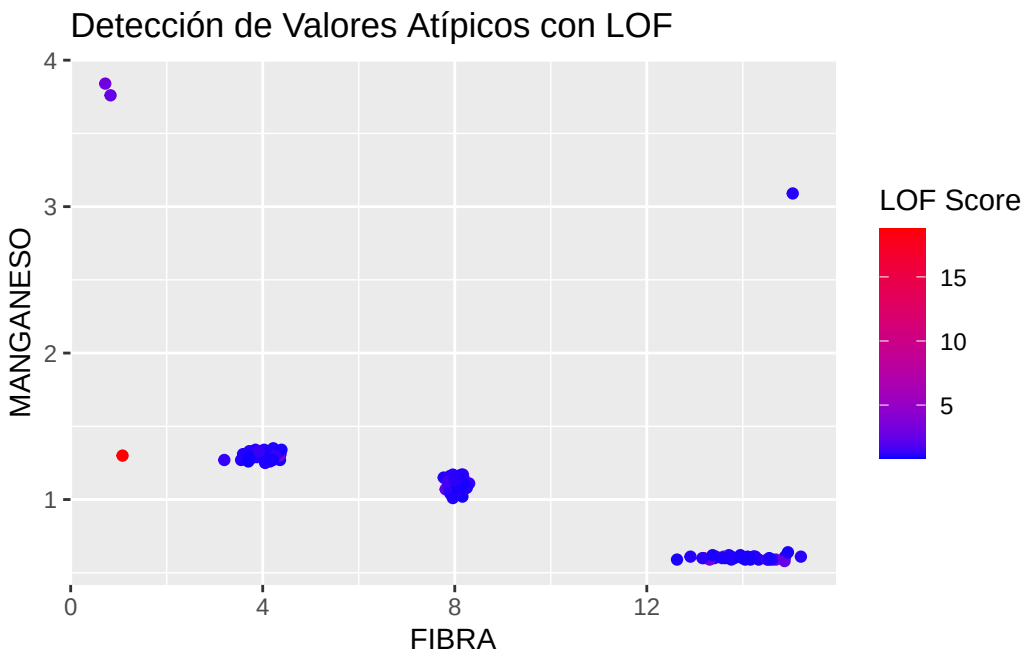
```
####Comprobamos las cuantitativas
ggplot(datos, aes(x = MANGANESO, y = SELENIO, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



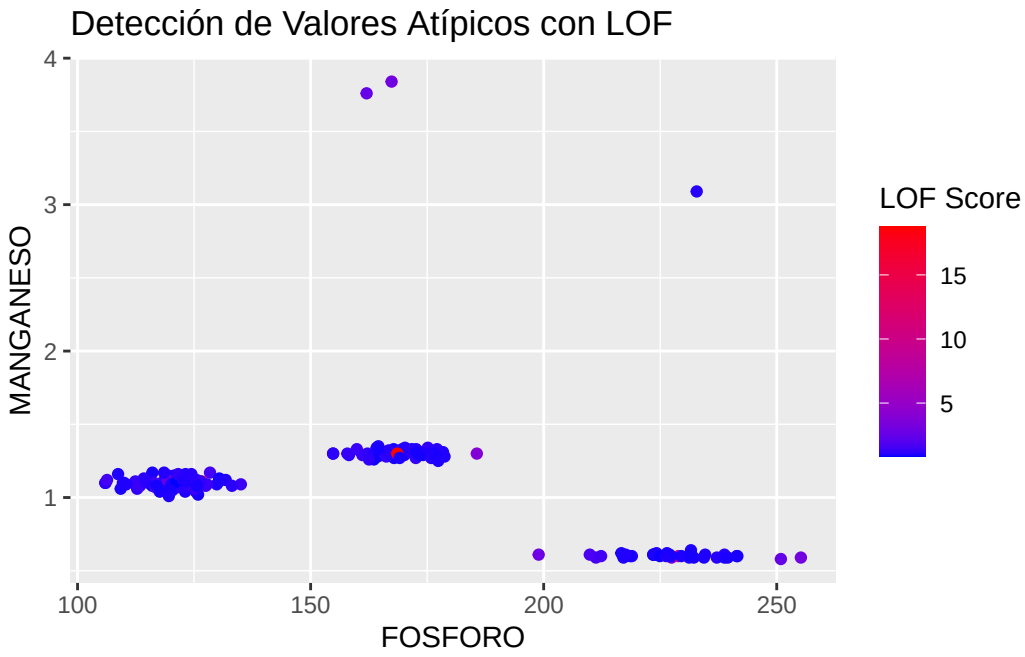

```
ggplot(datos, aes(x = CALORIAS, y = MANGANESO, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```




```
ggplot(datos, aes(x = FIBRA, y = MANGANESO, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```




```
ggplot(datos, aes(x = FOSFORO, y = MANGANESO, colour = lof)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score") +
  labs(title = "Detección de Valores Atípicos con LOF")
```



En el estudio multivariante vemos como hay 3 observaciones que tienen un LOF completamente elevado, estos corresponden a los outliers de la variable de SELENIO. No vemos en el resto de las variables que estas observaciones se comporten de forma rara, simplemente son tan atípicas en la variable SELENIO, que el algoritmo LOF las ha detectado. Sólo borraremos el valor de SELENIO.