Practica2

Silvia Pineda

Instrucciones (leer antes de empezar)

Modifica dentro del documento Practica2.qmd tus datos personales (nombre y DNI) ubicados en la cabecera del archivo.

Asegúrate, ANTES de seguir editando el documento, que el archivo .qmd se renderiza correctamente y se genera el .html correspondiente en tu carpeta local de tu ordenador.

Los chunks (cajas de código) creados están o vacíos o incompletos, de ahí que la mayoría tengan la opción #| **eval: false**. Una vez que edites lo que consideres, debes ir cambiando cada chunck a #| **eval: true** (o quitarlo directamente) para que se ejecuten.

ENUNCIADO DE LA PRÁCTICA

Para esta práctica vais a usar la base de datos **pacientes.csv** cuya definición encontraréis en el TEMA 2/ PRÁCTICA 2 en los apuntes de gitbook. Esta base de datos contiene información sobre pacientes que fueron sometidos a una intervención quirúrgica y se analizó si tuvieron alguna complicación durante la intervención. Para ello se tomaron varias mediciones cardiacas y alguna información más descrita en la tabla que se proporciona junto con los datos.

CARGA DE DATOS

```
datos<-read.csv("pacientes.csv")
library(ggplot2)
```

EJERCICIO 1: ¿El tipo de variable está declarado de forma correcta? Modifícalo si no es así

```
str(datos)
```

```
'data.frame':   60 obs. of  9 variables:
 $ Obs  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Patno: chr  "1" "2" "3" "4" ...
 $ Sex  : chr  "M" "F" "X" "F" ...
 $ Visit: chr  "11/11/98" "11/13/98" "10/21/98" "1/1/99" ...
 $ HR   : int  89 69 79 81 79 74 90 74 95 74 ...
 $ SBP  : int  119 110 111 118 118 122 129 113 131 117 ...
 $ DBP  : int  98 81 83 93 85 94 93 90 99 83 ...
 $ Dx   : chr  "1" "X" "3" "4" ...
 $ AE   : chr  "1" "0" "0" "A" ...
```

```
datos[,c("Sex","AE","Dx")]<-
    lapply(datos[,c("Sex","AE","Dx")],as.factor)

datos$Visit<-as.Date(datos$Visit,format="%m/%d/%y")
str(datos)
```

```
'data.frame':   60 obs. of  9 variables:
 $ Obs  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Patno: chr  "1" "2" "3" "4" ...
 $ Sex  : Factor w/ 6 levels "","2","f","F",..: 5 4 6 4 5 1 5 5 4 5 ...
 $ Visit: Date, format: "1998-11-11" "1998-11-13" ...
 $ HR   : int  89 69 79 81 79 74 90 74 95 74 ...
 $ SBP  : int  119 110 111 118 118 122 129 113 131 117 ...
 $ DBP  : int  98 81 83 93 85 94 93 90 99 83 ...
 $ Dx   : Factor w/ 6 levels "0","1","2","3",..: 2 6 4 5 2 4 1 1 3 5 ...
 $ AE   : Factor w/ 3 levels "0","1","A": 2 1 1 3 2 1 2 2 2 2 ...
```

```
#datos$Patno<-as.numeric(datos$Patno)
#which(is.na(datos$Patno)==T)
```

Las variables Sex, AE y Dx son variables cualitativas y las declaramos como factor.

La variable Visit es una fecha y la declaramos como fecha.

La variable Patno es un identificador numérico, pero al pasarla a numeric, nos salta un warning porque hay datos erróneos que debemos corregir.

EJERCICIO 2: ¿Todas las variables toman valores correctos? Si hay valores incorrectos haz que aparezcan como missing (NA) o si los puedes identificar, corrígelos

```
summary(datos)
```

```
      Obs            Patno              Sex         Visit           
 Min.   : 1.00   Length:60          1    : 1   Min.   :1998-03-28  
 1st Qu.:15.75   Class :character   2    : 1   1st Qu.:1998-07-27  
 Median :30.50   Mode  :character   f    : 2   Median :1999-01-01  
 Mean   :30.50                      F    :27   Mean   :1999-01-22  
 3rd Qu.:45.25                      M    :28   3rd Qu.:1999-07-07  
 Max.   :60.00                      X    : 1   Max.   :1999-11-12  
                                               NA's   :10          
       HR              SBP              DBP          Dx      AE     
 Min.   : 50.00   Min.   : 60.0   Min.   : 40.00   0:10   0:28  
 1st Qu.: 68.50   1st Qu.:105.0   1st Qu.: 83.00   1:18   1:31  
 Median : 75.00   Median :113.0   Median : 89.00   2: 6   A: 1  
 Mean   : 83.17   Mean   :110.4   Mean   : 88.77   3:11          
 3rd Qu.: 82.00   3rd Qu.:117.2   3rd Qu.: 93.00   4:13          
 Max.   :500.00   Max.   :132.0   Max.   :130.00   X: 2          
```

```
library(car)
```

```
Loading required package: carData
```

```
datos$Patno[5]<-5
datos$Patno<-as.numeric(datos$Patno)
datos$Patno<-replace(datos$Patno,datos$Patno==121,21)


datos$Sex<-car::recode(datos$Sex,"'f'='F';'2'='F';'X' = NA ;'' = NA")
table(datos$Sex)
```

```
 F  M 
30 28 
```

```r
datos$AE<-car::recode(datos$AE,"'A' = 0")
table(datos$AE)
```

```
 0  1
29 31
```

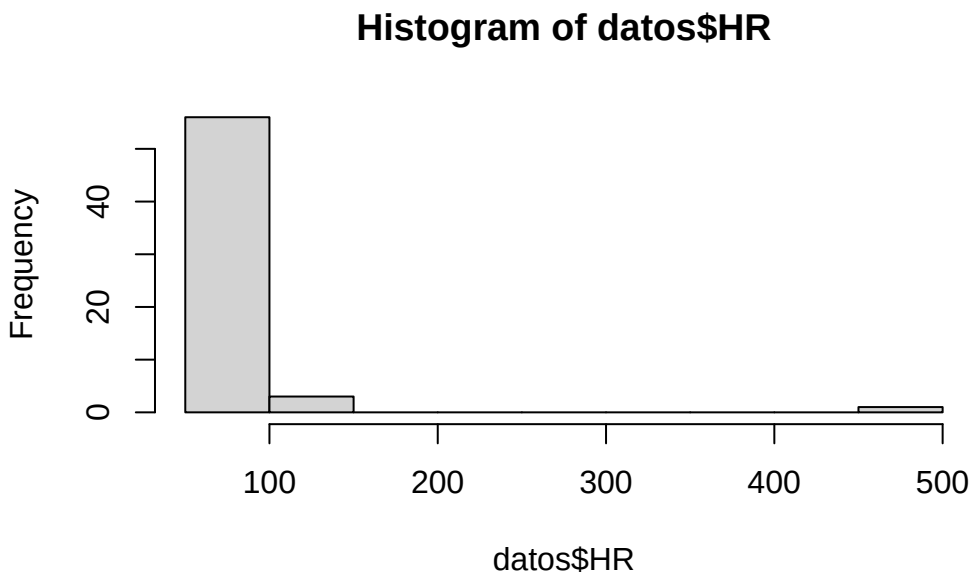
```r
datos$Dx<-car::recode(datos$Dx,"'X'=NA")
table(datos$Dx)
```

```
 0  1  2  3  4
10 18  6 11 13
```

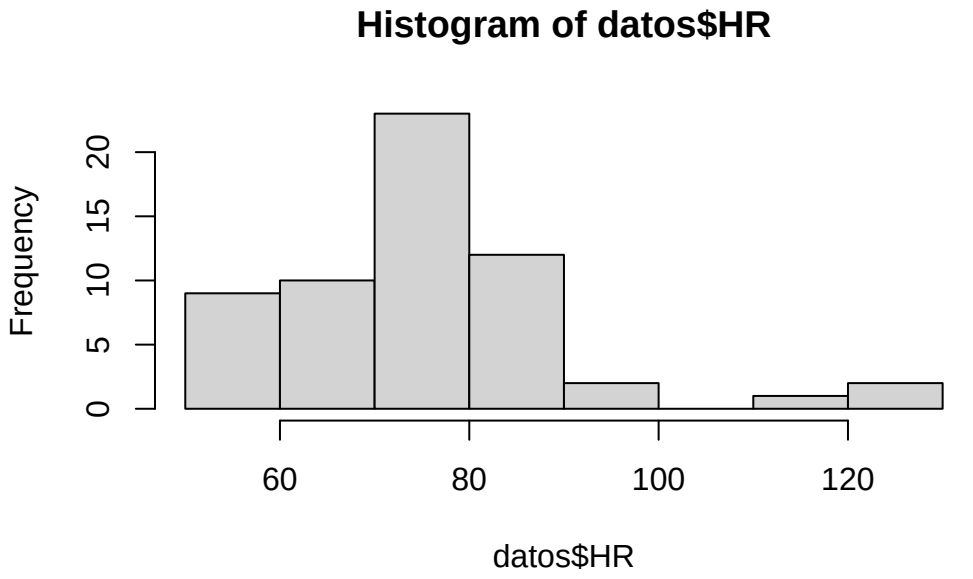
```r
summary(datos)
```

```
      Obs             Patno          Sex         Visit
 Min.   : 1.00   Min.   : 1.00   F   :30   Min.   :1998-03-28
 1st Qu.:15.75   1st Qu.:15.75   M   :28   1st Qu.:1998-07-27
 Median :30.50   Median :30.50   NA's: 2   Median :1999-01-01
 Mean   :30.50   Mean   :30.50             Mean   :1999-01-22
 3rd Qu.:45.25   3rd Qu.:45.25             3rd Qu.:1999-07-07
 Max.   :60.00   Max.   :60.00             Max.   :1999-11-12
                                           NA's   :10
       HR              SBP             DBP             Dx      AE
 Min.   : 50.00   Min.   : 60.0   Min.   : 40.00   0   :10   0:29
 1st Qu.: 68.50   1st Qu.:105.0   1st Qu.: 83.00   1   :18   1:31
 Median : 75.00   Median :113.0   Median : 89.00   2   : 6
 Mean   : 83.17   Mean   :110.4   Mean   : 88.77   3   :11
 3rd Qu.: 82.00   3rd Qu.:117.2   3rd Qu.: 93.00   4   :13
 Max.   :500.00   Max.   :132.0   Max.   :130.00   NA's: 2
```

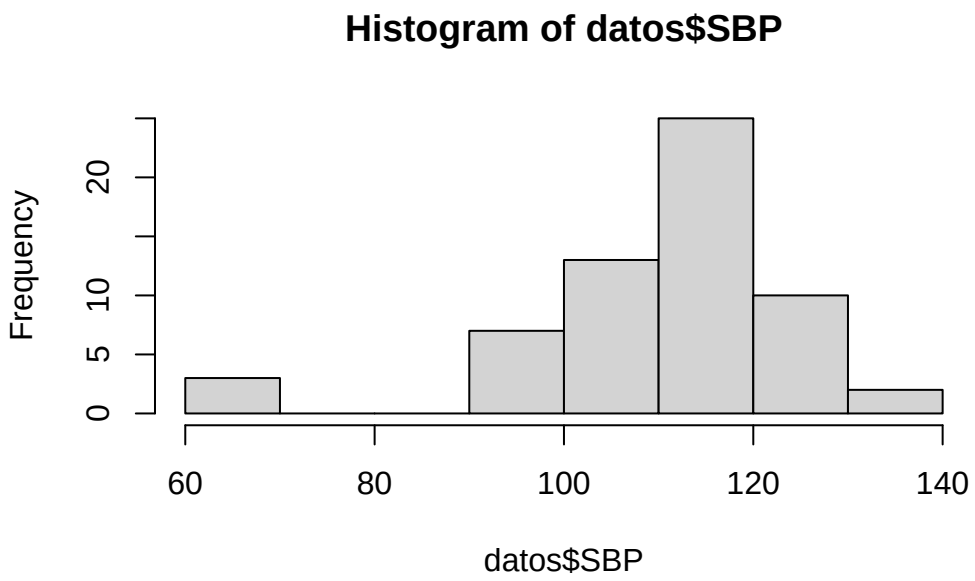
```r
hist(datos$HR)
```



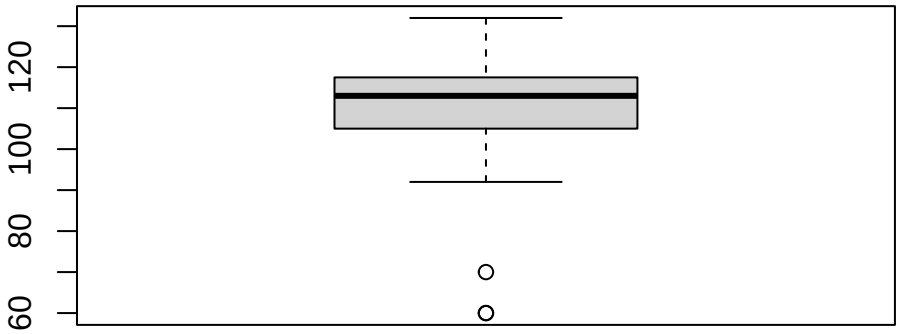

```
datos$HR<-replace(datos$HR,datos$HR==500,NA)
hist(datos$HR)
```



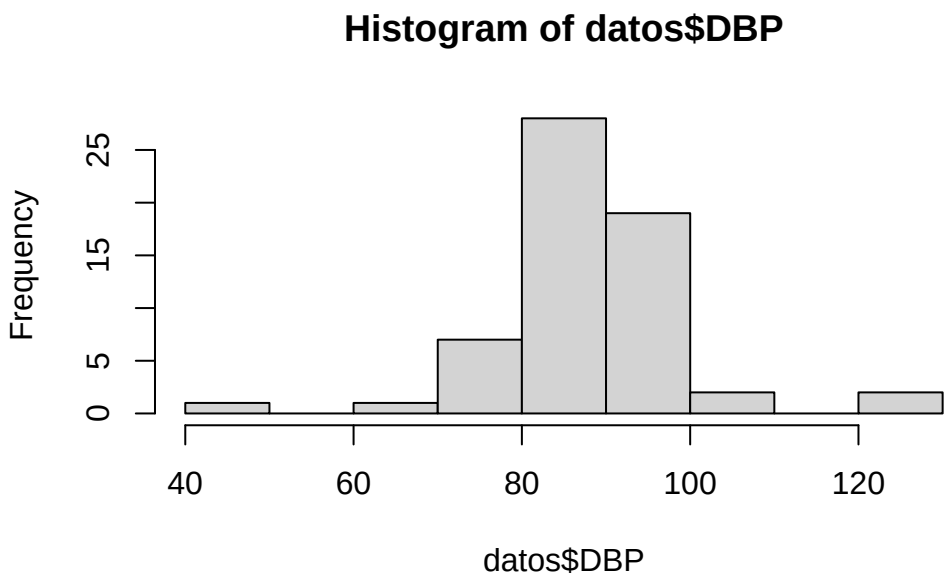
```r
hist(datos$SBP)
```



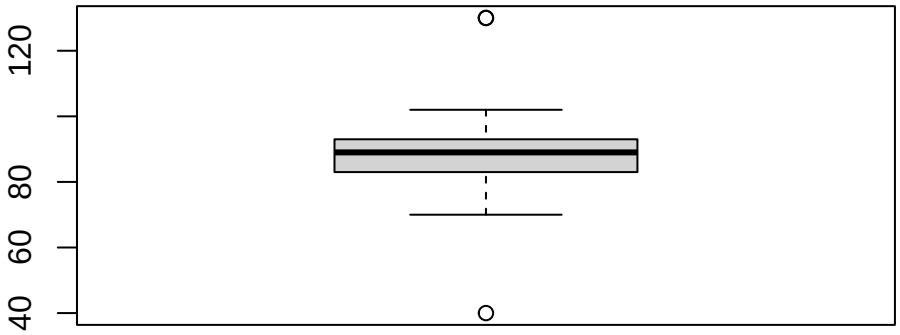

```r
boxplot(datos$SBP)
```



```r
hist(datos$DBP)
```




```r
boxplot(datos$DBP)
```



Patno tiene un valor xx5 que parece un error que podemos rescatar ya que es la observación 5 y lo mismo con la observación 21 que pone 121 y es un error que podemos convertir en 21.

En Sex podemos rescatar la f y el 2 a F y el resto lo ponemos como NA.

En AE la A podemos interpretar que es Ausencia y rescatarla como un 0.

En Dx el X no sabemos que puede ser pero la definición es que los diagnósticos son numéricos, por tanto lo pondremos como NA.

El valor 500 de HR es un error porque es imposible.

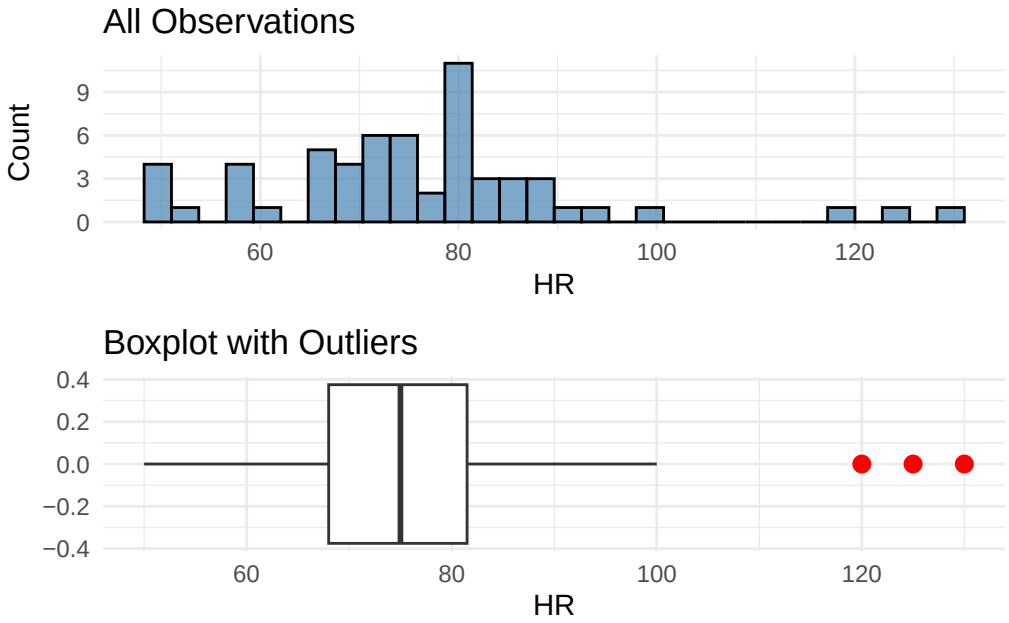
EJERCICIO 3: Realiza un estudio de las variables numéricas para ver si hay datos atípicos y eliminalos o no según consideres. Haz el estudio univariante y bivariante.

```
###Univariante
source("outliers.R")
```

```
Attaching package: 'dplyr'

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

    recode

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

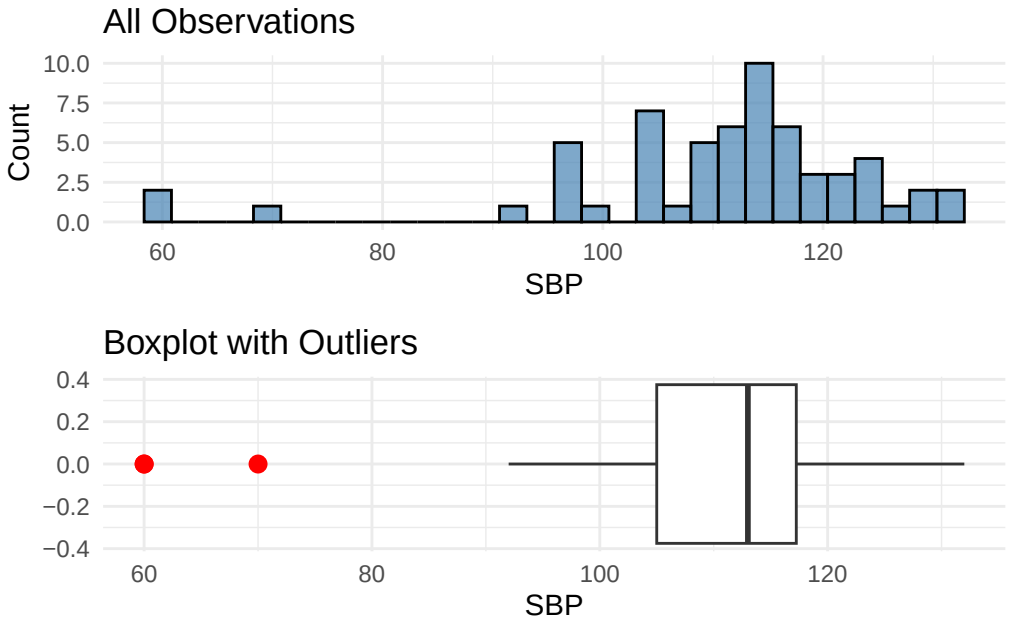
```
outliers(datos,"HR")
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_bin()`).

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```




```
 Outliers identified in HR :  3 outliers
 Proportion (%) of outliers: 5.08 %

[1] 125 120 130
```

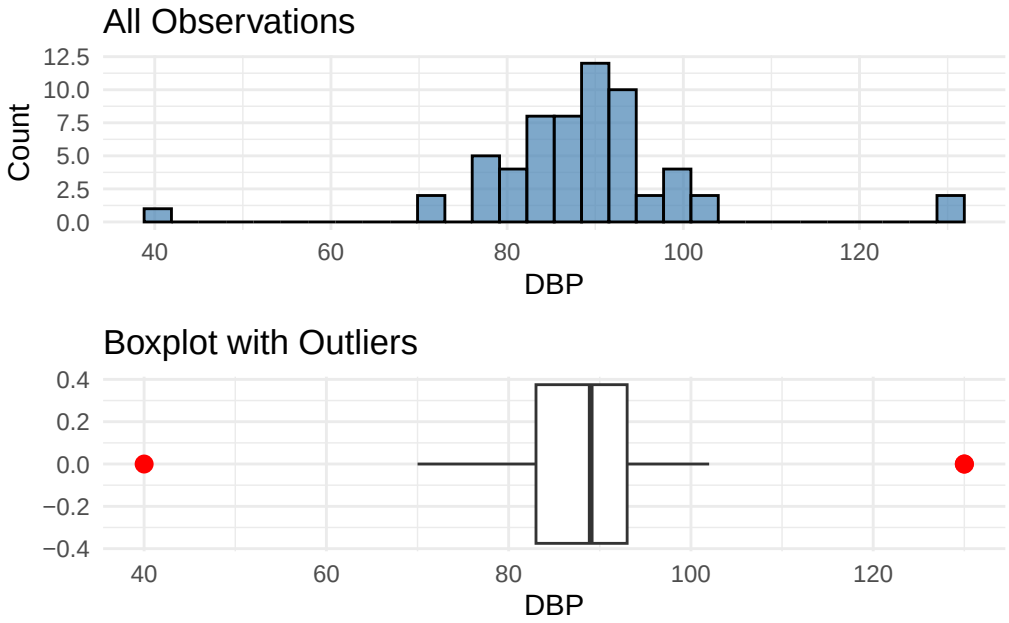
```
outliers(datos,"SBP")
```




```
 Outliers identified in SBP :  3 outliers
 Proportion (%) of outliers: 5 %
```

```
[1] 60 70 60
```

```
outliers(datos,"DBP")
```

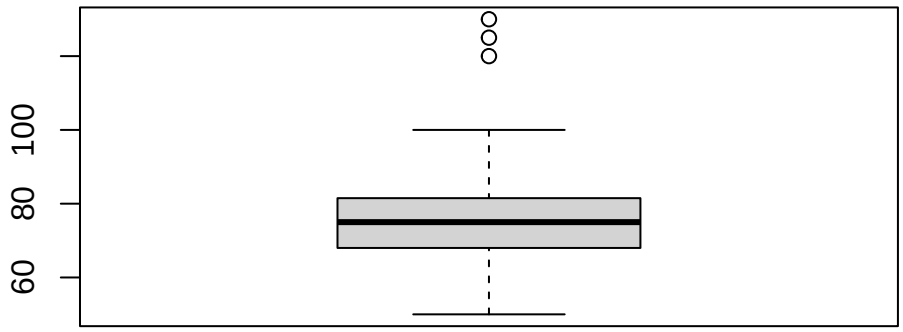



```
Outliers identified in DBP :  3 outliers
Proportion (%) of outliers: 5 %
```

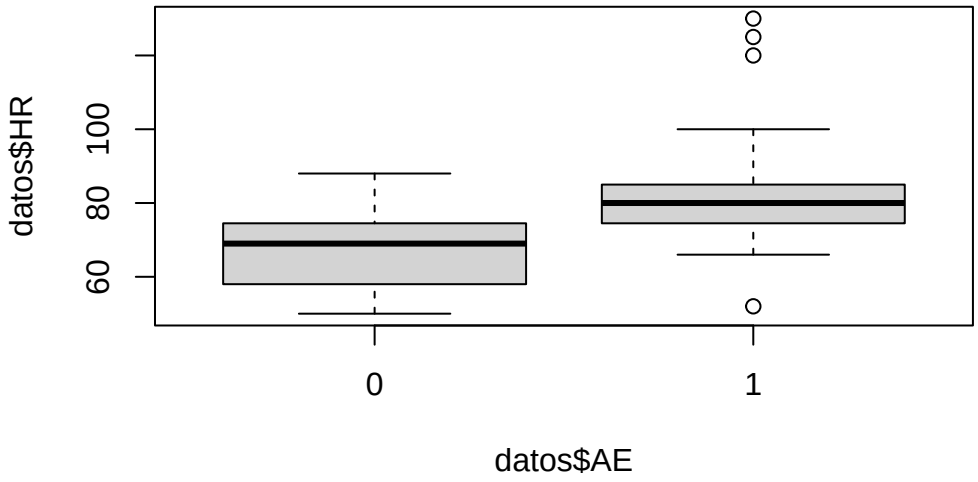
```
[1] 130  40 130
```

Las tres variables numéricas HR, SBP y DBP tienen outliers, el porcentaje es alrededor del 5% para las 3 variables y unos puntos bastante claros que salen de los boxplots. En SBP y DBP no parecen que sean por distribuciones muy asimétricas, pero en HR podríamos pensar que son unos casos un poco extremos pero quizás no atícpicos, veremos que pasa en el bivariante.

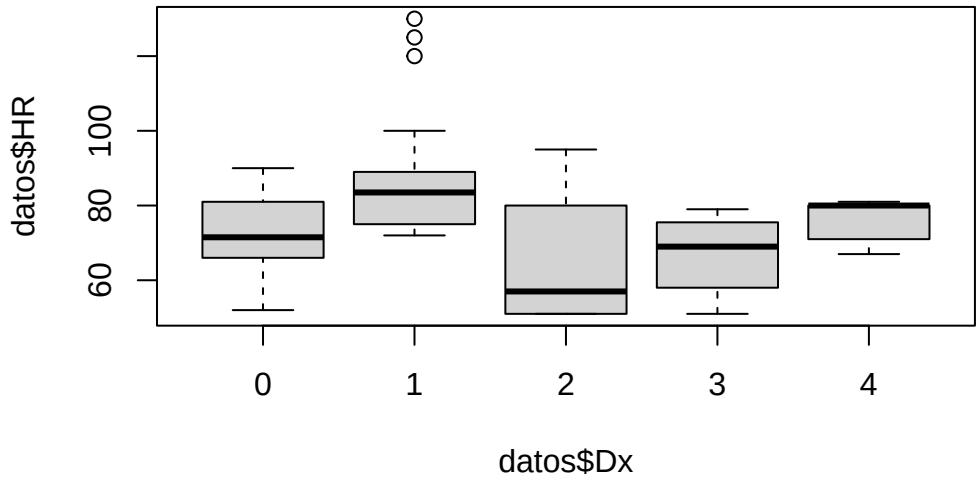
```
boxplot(datos$HR)
```




```
boxplot(datos$HR ~ datos$AE)
```

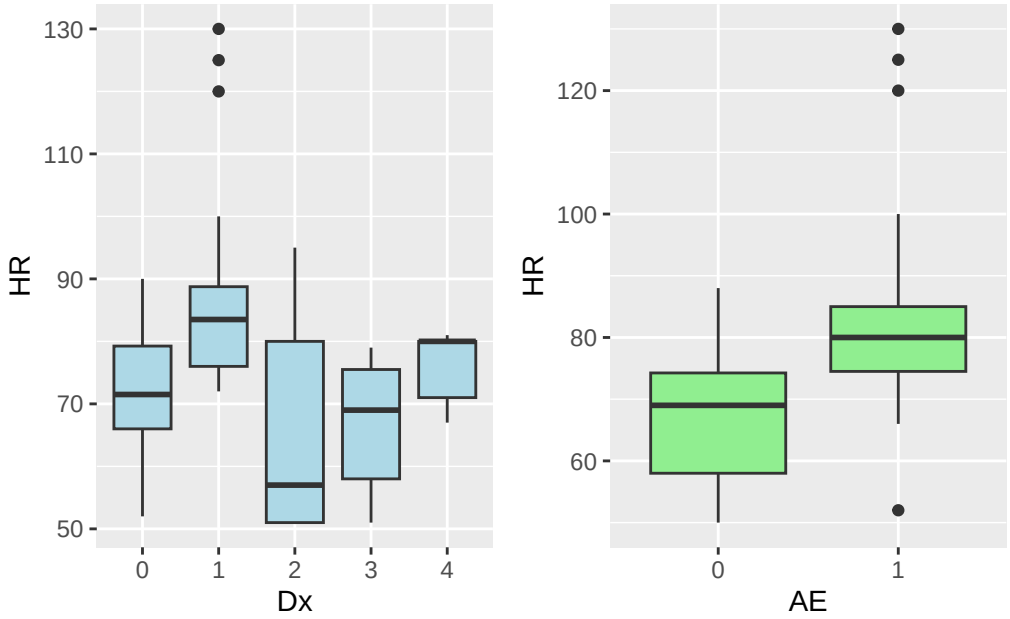


```
boxplot(datos$HR ~ datos$Dx)
```

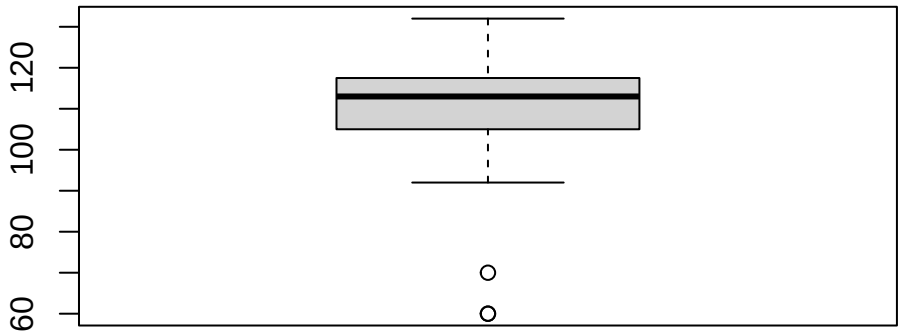



```
gp1<-ggplot(datos %>% filter(!is.na(Dx)), aes(x = Dx, y = HR)) +
  geom_boxplot(fill = "lightblue")
gp2<-ggplot(datos, aes(x = AE, y = HR)) +
  geom_boxplot(fill = "lightgreen")
gp1+gp2
```

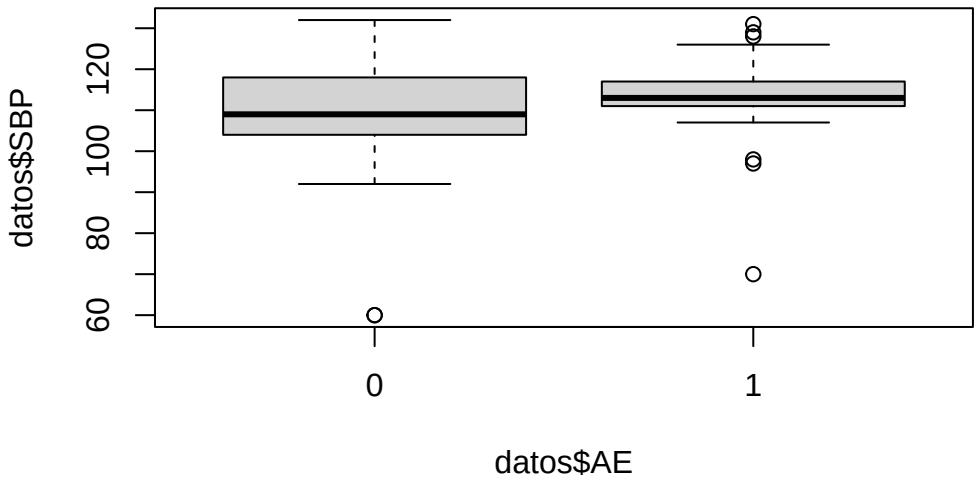
```
Warning: Removed 1 row containing non-finite outside the scale range (`stat_boxplot()`).
Removed 1 row containing non-finite outside the scale range (`stat_boxplot()`).
```



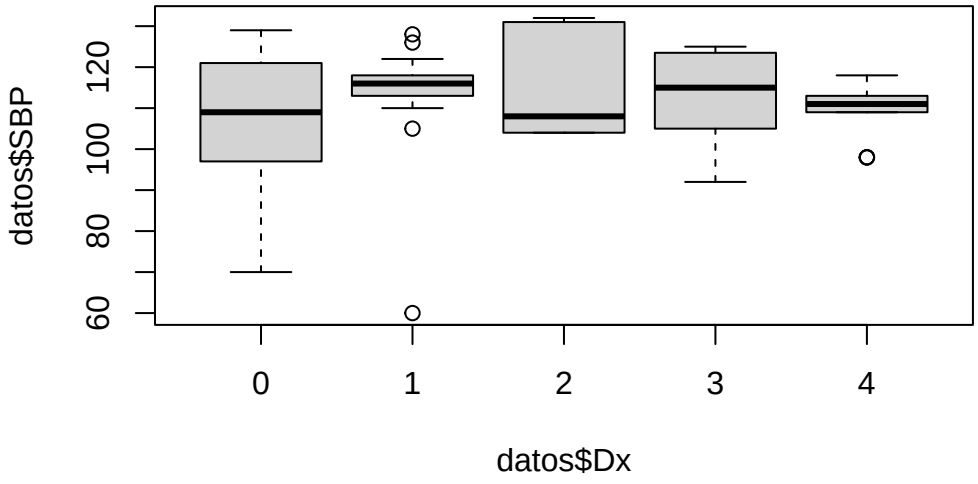

```
###No son outliers

boxplot(datos$SBP)
```



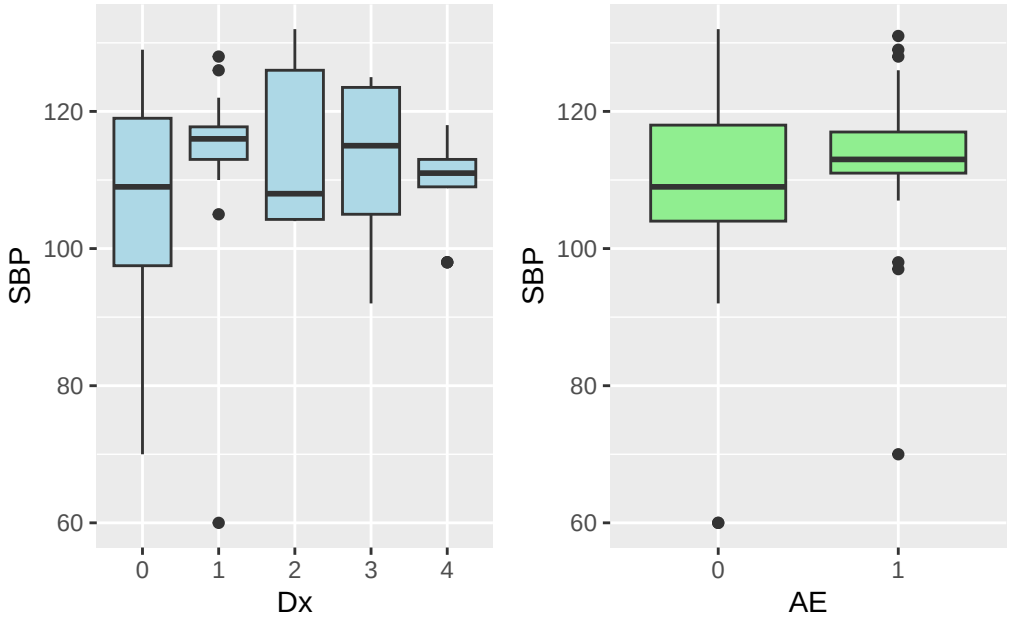
```
boxplot(datos$SBP ~ datos$AE)
```



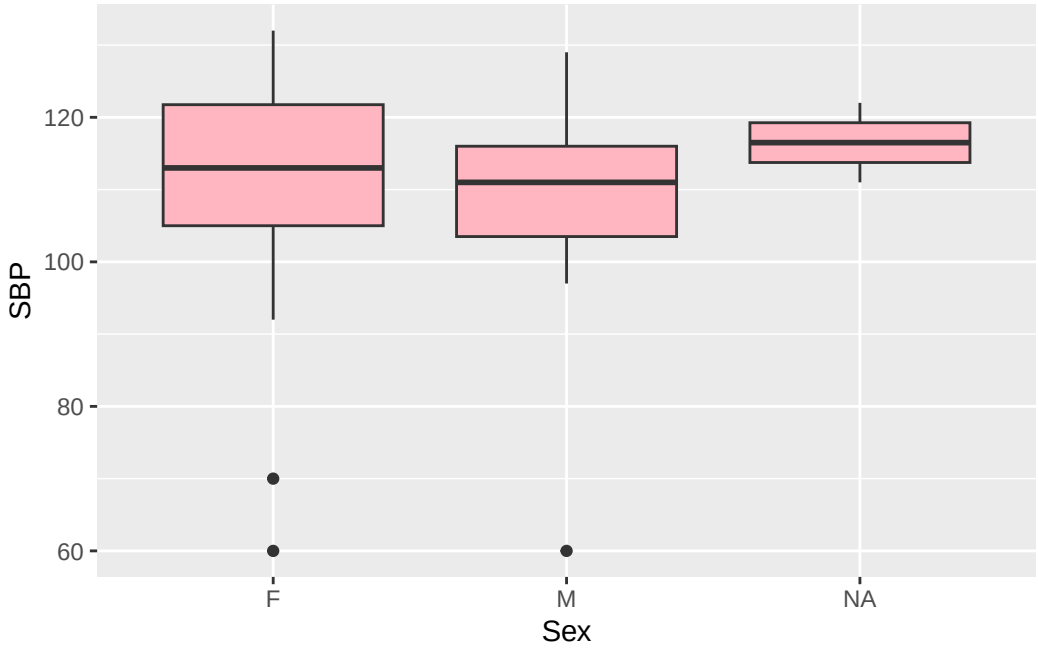

```
boxplot(datos$SBP ~ datos$Dx)
```



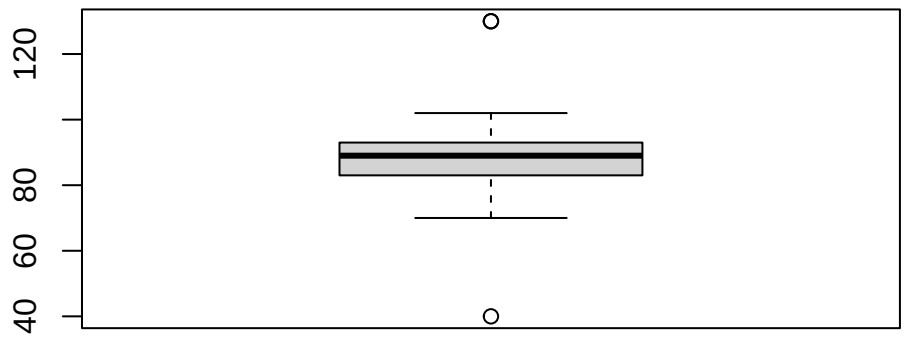
```
gp1<-ggplot(datos %>% filter(!is.na(Dx)), aes(x = Dx, y = SBP)) +
  geom_boxplot(fill = "lightblue")
gp2<-ggplot(datos, aes(x = AE, y = SBP)) +
  geom_boxplot(fill = "lightgreen")
gp1+gp2
```



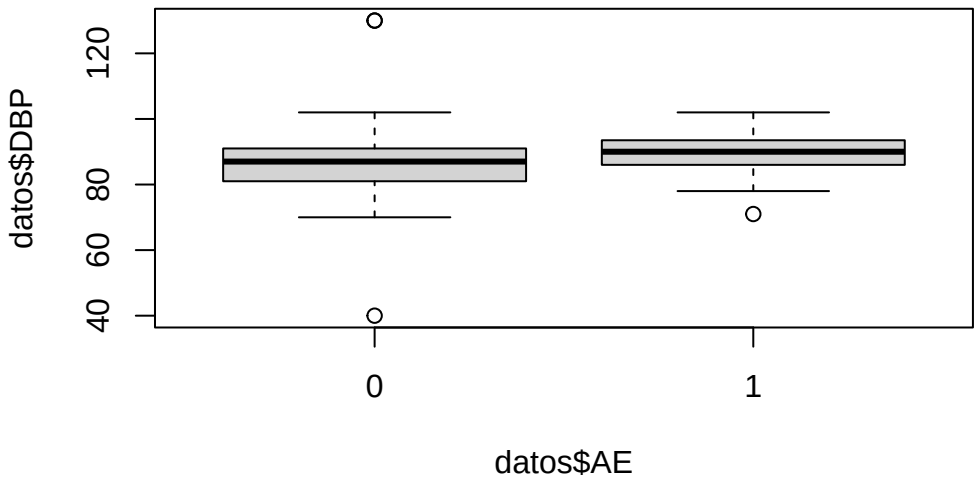

```
ggplot(datos, aes(x = Sex, y = SBP)) +
  geom_boxplot(fill = "lightpink")
```



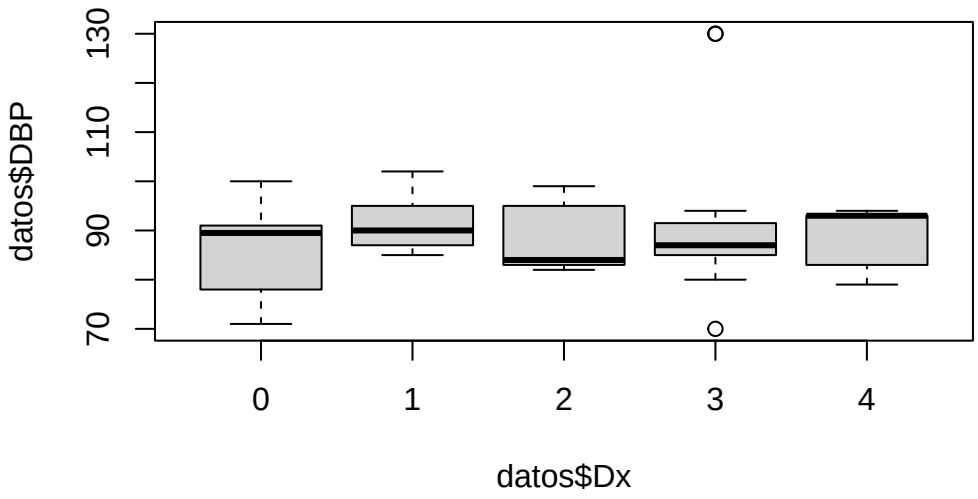

```
##Estos se borran


boxplot(datos$DBP)
```



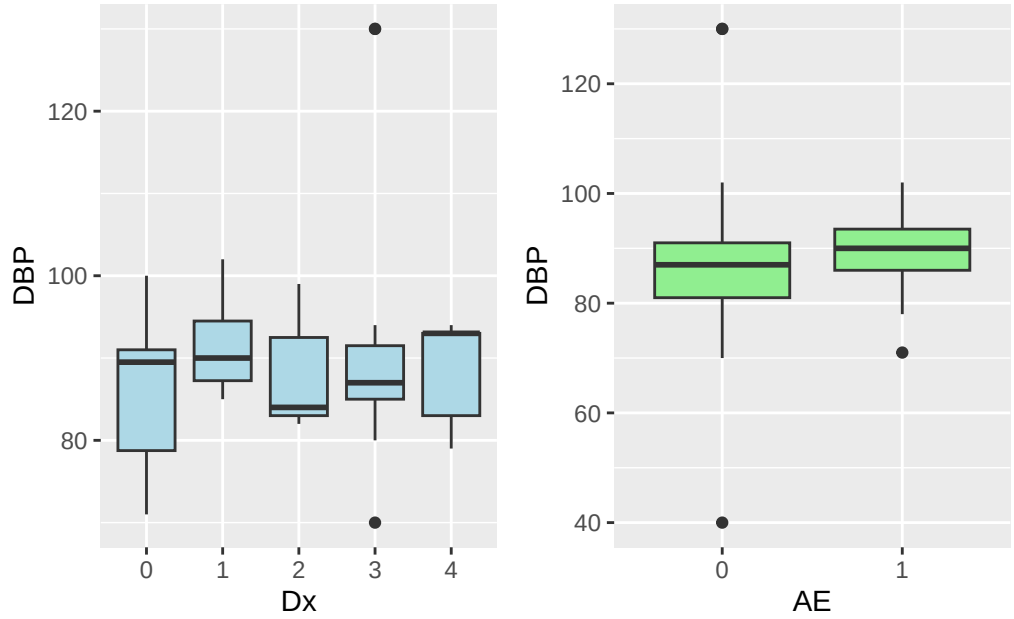

```
boxplot(datos$DBP ~ datos$AE)
```



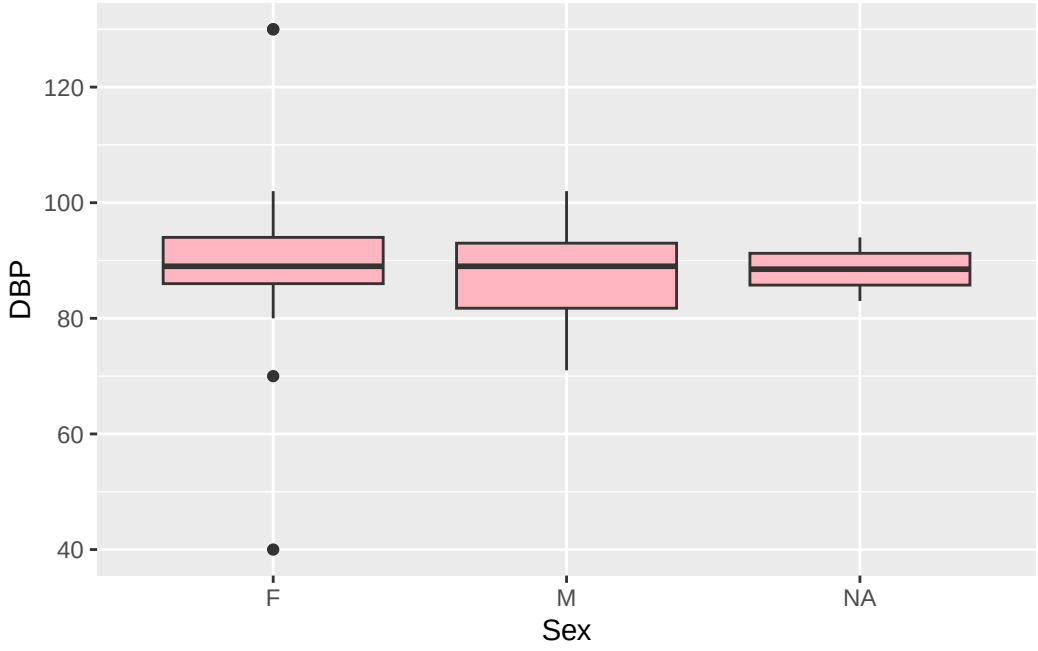
```r
boxplot(datos$DBP ~ datos$Dx)
```



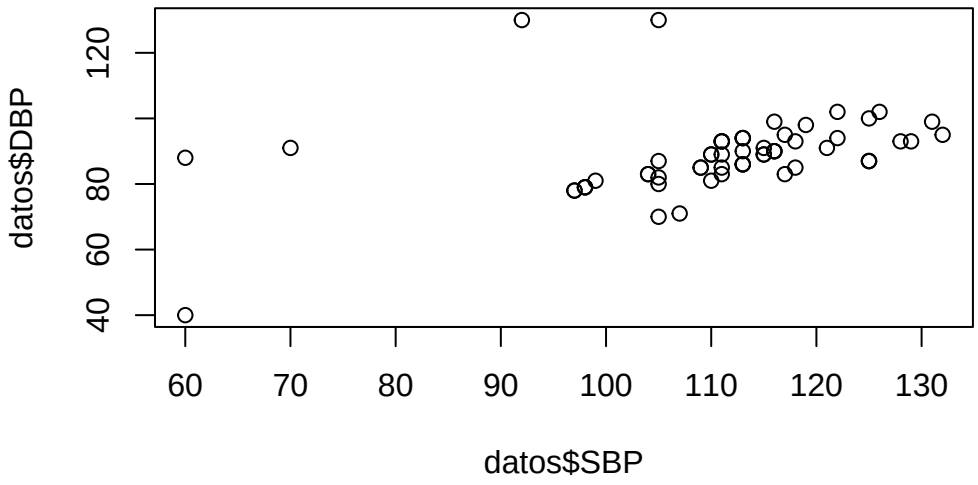

```r
##Estos se borran

gp1<-ggplot(datos %>% filter(!is.na(Dx)), aes(x = Dx, y = DBP)) +
  geom_boxplot(fill = "lightblue")
gp2<-ggplot(datos, aes(x = AE, y = DBP)) +
  geom_boxplot(fill = "lightgreen")
gp1+gp2
```



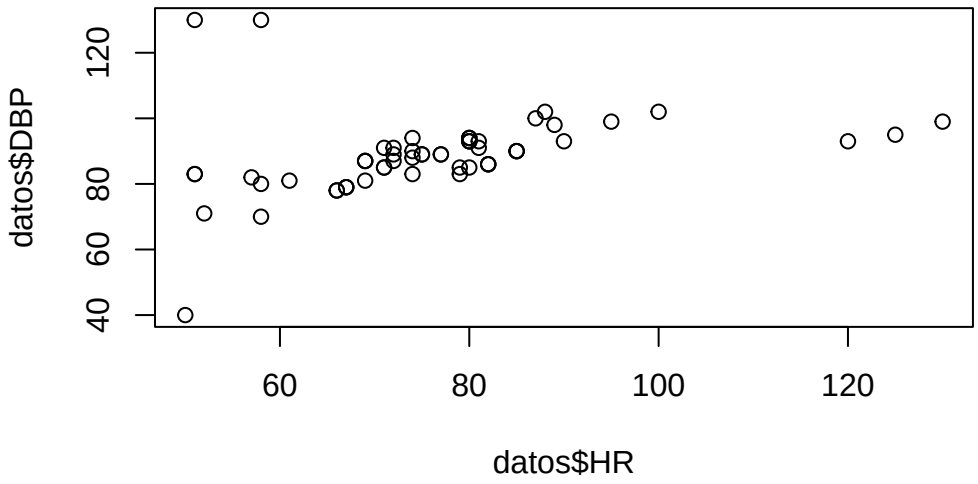

```
ggplot(datos, aes(x = Sex, y = DBP)) +
  geom_boxplot(fill = "lightpink")
```



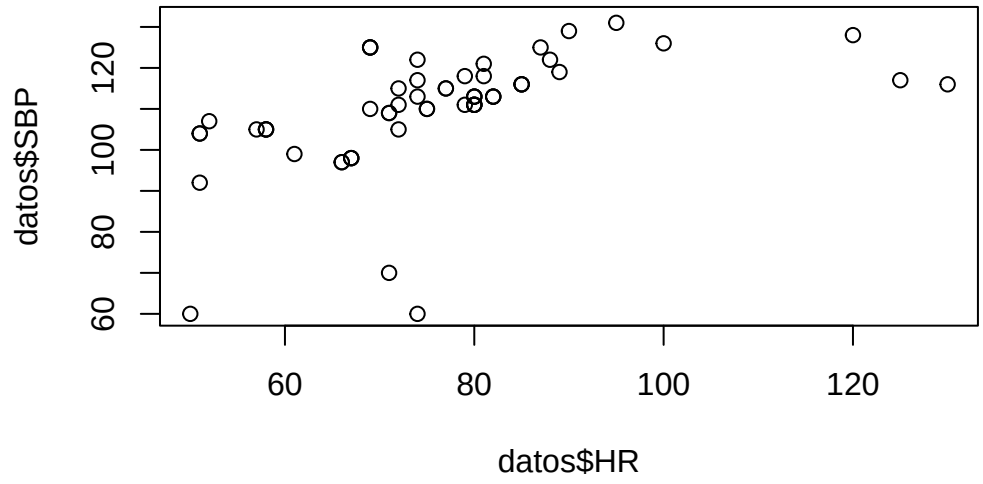
```r
plot(datos$SBP,datos$DBP)
```



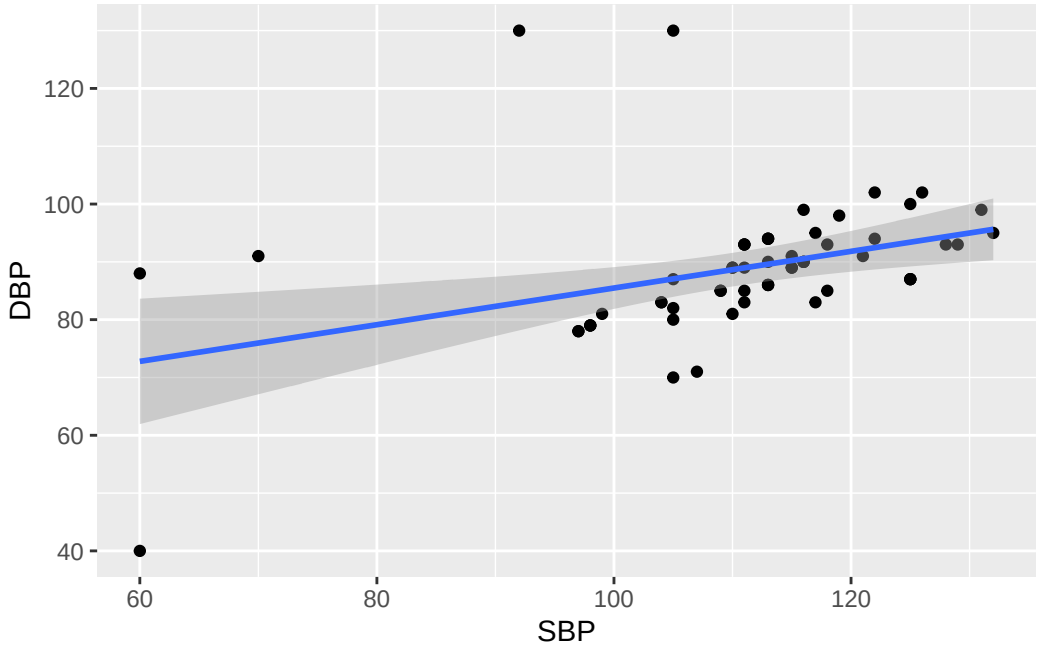

```r
plot(datos$HR,datos$DBP)
```



```r
plot(datos$HR,datos$SBP)
```



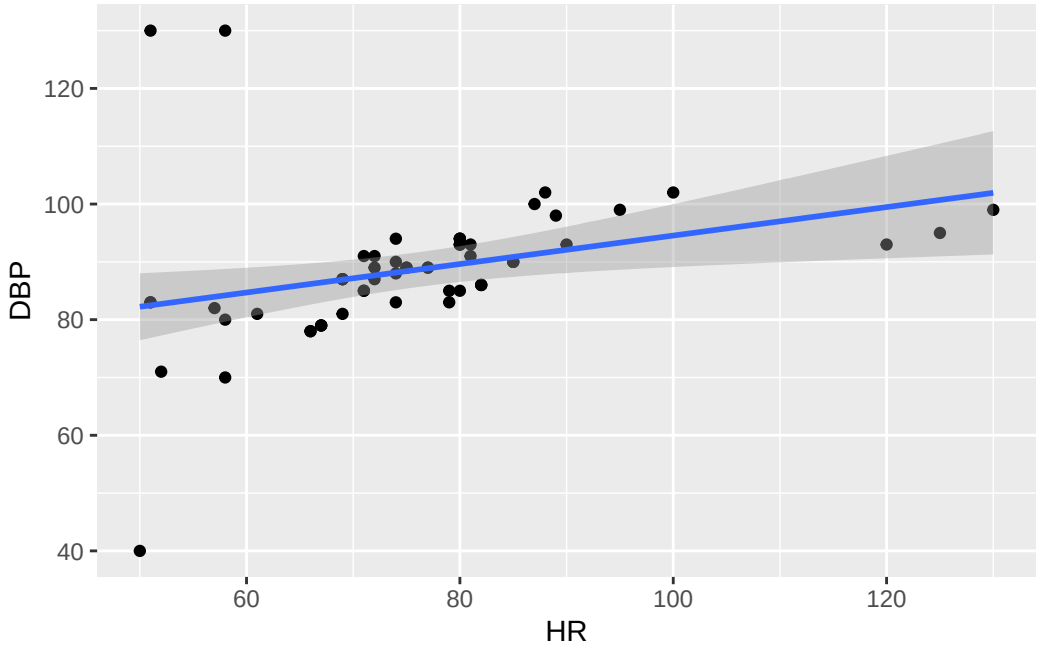

```r
ggp <- ggplot(datos,aes(SBP, DBP)) + geom_point()
ggp + stat_smooth(method = "lm",
                  formula = y ~ x,
                  geom = "smooth")
```




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

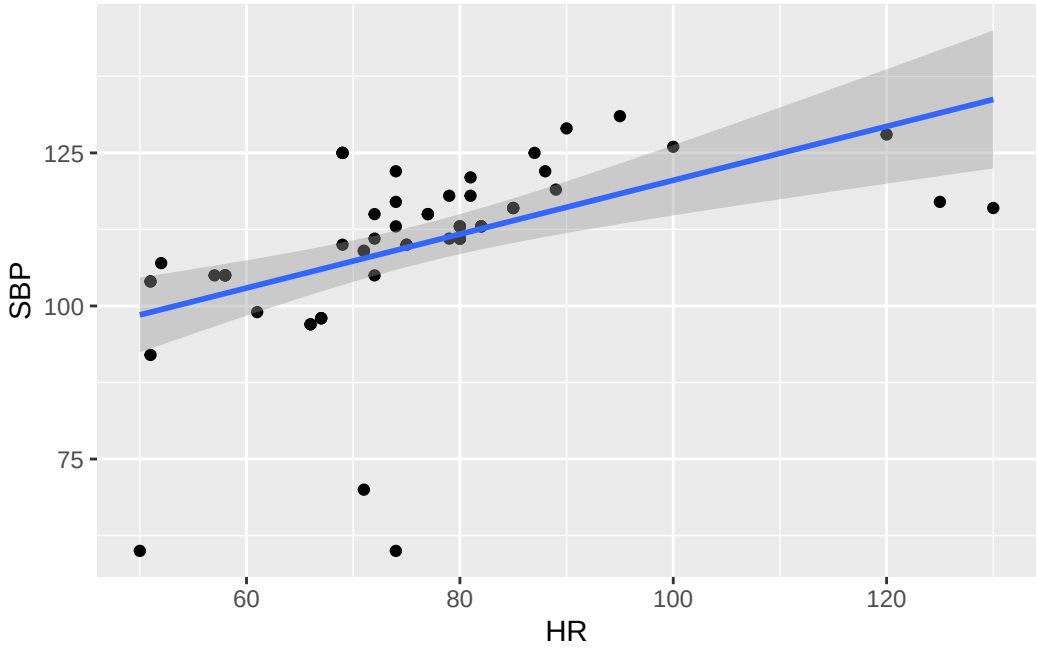
```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```




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

```
Warning: Removed 1 row containing non-finite outside the scale range (`stat_smooth()`).
Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



Para los outliers del HR vemos que pertenecen a aquellos individuos que tuvieron un eveno adverso y que corresponden al diagnóstico 1 que es un diagnóstico que corresponde a tener un mayor HR, por tanto son parte de una asociación y no los borraremos.

En el caso de SBP los outliers que corresponden a individuos con una muy baja presión sistólica no tienen una asociación con ninguna de las dos variables objetivo, tampoco con sexo, por tanto los borraremos.

En el caso de DBP, los outliers que pertencen a dos altas y una baja, tampoco parece que tengan nignuna relación, por tanto las borraremos.

Si quedan dudas, los gráficos de las variables continuas nos muestran que los outliers de DBP y SBP quedan completamente fuera de la nube de puntos, mientras que los de HR quedan alejados pero no se van demasiado.

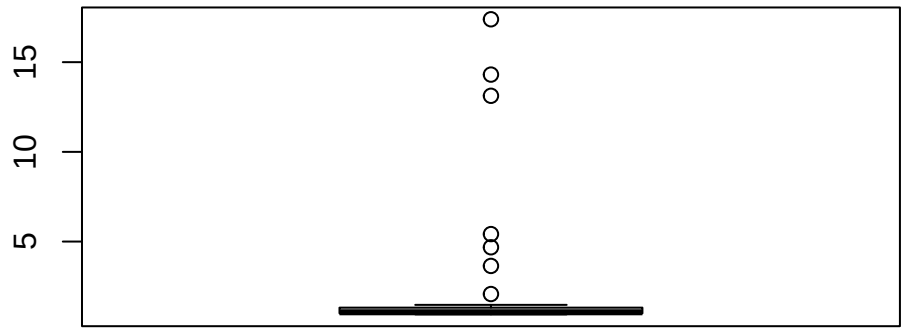
EJERCICIO 4: Aplica el algoritmo LOF para hacer un estudio multivariante de detección de datos atípicos y comenta que observan en los resultados.

```
library(dbscan)
```

```
Attaching package: 'dbscan'
```

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

    as.dendrogram
```

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

datosLOF<-datos[complete.cases(datos[,c("HR","SBP","DBP")]),]

# Aplica LOF para detección de valores atípicos
k<-round(log(nrow(datos)))
lof_resultados <- lof(datosLOF[,c("HR","SBP","DBP")],minPts = k)

### Los que tienen un LOF mayor de 1.5
boxplot(lof_resultados)
```

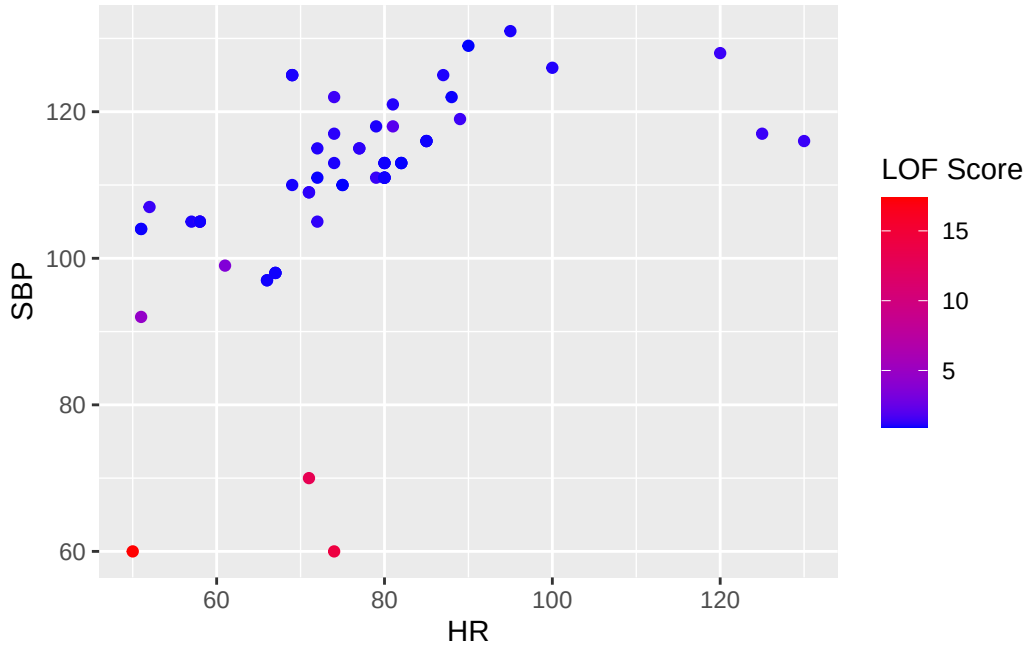



```r
datosLOF$lof<-lof_resultados
datosLOF[which(datosLOF$lof>2),]
```

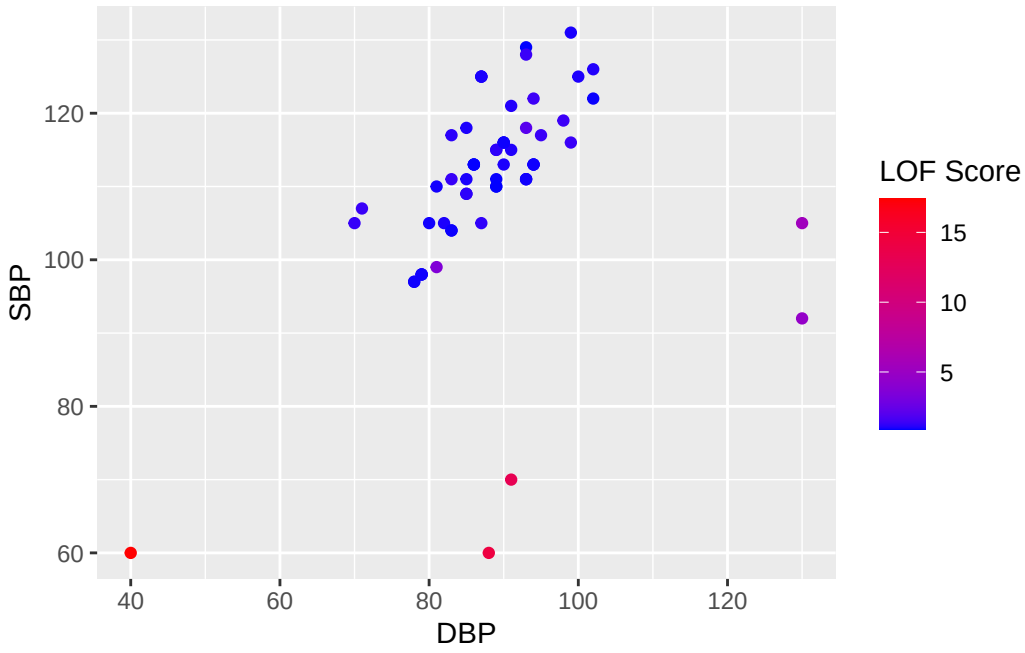
```
   Obs Patno Sex      Visit HR SBP DBP   Dx AE       lof
4    4     4   F 1999-01-01 81 118  93    4  0  2.074490
13  13    13   M 1998-10-12 61  99  81    0  0  3.641718
```

```
15  15    15   M 1999-02-02 74  60  88    1  0 14.307277
18  18    18   F        <NA> 51  92 130    3  0  4.681254
25  25    25   F 1998-12-31 71  70  91    0  1 13.125577
29  29    29   F 1998-03-28 50  60  40 <NA>  0 17.390979
35  35    35   F 1999-11-12 58 105 130    3  0  5.416172
```

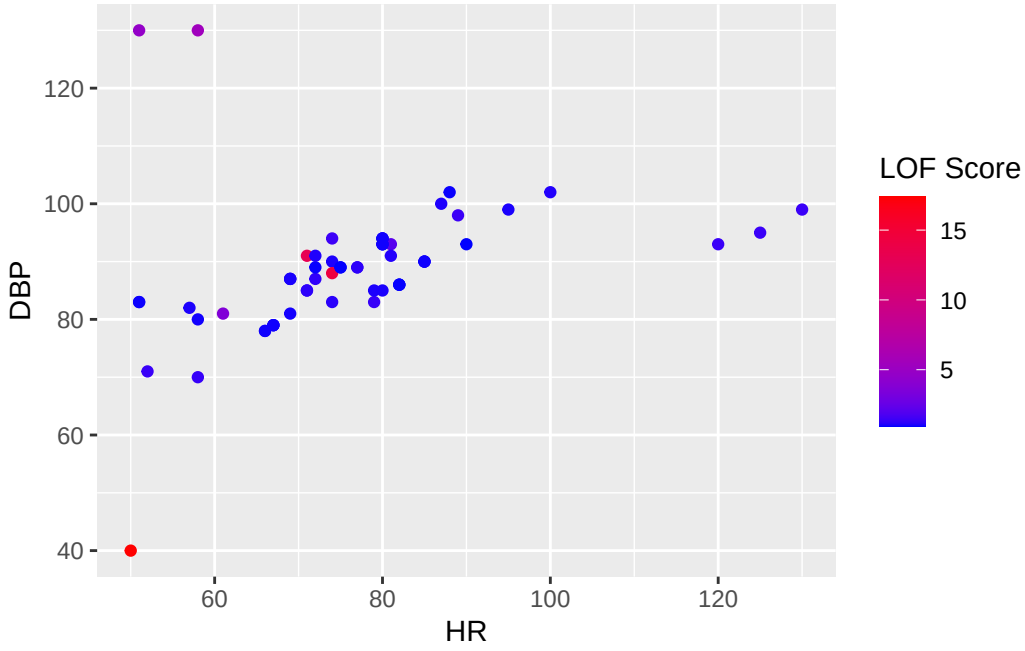
```
ggplot(datosLOF, aes(x = HR, y = SBP, colour = lof_resultados)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score")
```




```
ggplot(datosLOF, aes(x = DBP, y = SBP, colour = lof_resultados)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score")
```




```
ggplot(datosLOF, aes(x = HR, y = DBP, colour = lof_resultados)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red", name = "LOF Score")
```



Si miramos los datos que tienen un lof>2 vemos que los individuos con un lof muy muy alto corresponden a los 3 outliers de la variable SBP. Los otros dos a los outliers de DBP. Además el individuo con el lof más elevado (17.39) corresponde al que tiene SBP y DBP como valor atípico y que el HR es extremadamente bajo, en este caso podemos contemplar borrar la observación entera. Y además ocurre una cosa interesante y es que aquellos individuos que tienen un lof elevado por el atípico de SBP o DBP, ocurre algo que es imposible y es que la presión diastólica (baja) es más elavada que la sistólica (alta) y esto es imposible.

CONCLUSIÓN FINAL:

Borrar los outliers de SBP y DBP, dejar los de HR porque no son outliers ya que se asocian con las variables objetivo y podríamos pensar en borrar la observación entera con el LOF más elevado que corresponde a la observación 29.

```
#Borramos los outliers de SBP y DBP

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

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

datos<-datos[-29,]
```