Ejercicios 1

Alberto Armijo Ruiz

8 de noviembre de 2018

1. Exploratory Data Analysis

a. Ejemplo 1, hip dataset

- Descargate el dataset hip con el siguiente commando
  hip <-read.table("http://astrostatistics.psu.edu/datasets/HIP_star.dat", header=T,fill=T)

```
hip  <-read.table("http://astrostatistics.psu.edu/datasets/HIP_star.dat", header=T,fill=T)
```

- Una vez descargado comprueba la dimensión y los nombres de las columnas del dataset. ¿Qué dimensión tiene? ¿qué datos alberga?

```
dim(hip)
```

```
## [1] 2719    9
```

```
colnames(hip)
```

```
## [1] "HIP"   "Vmag"  "RA"    "DE"    "Plx"   "pmRA"  "pmDE"  "e_Plx" "B.V"
```

```
str(hip)
```

```
## 'data.frame':    2719 obs. of  9 variables:
##  $ HIP  : int  2 38 47 54 74 81 110 135 143 149 ...
##  $ Vmag : num  9.27 8.65 10.78 10.57 9.93 ...
##  $ RA   : num  0.0038 0.111 0.1352 0.1517 0.2219 ...
##  $ DE   : num  -19.5 -79.1 -56.8 18 35.8 ...
##  $ Plx  : num  21.9 23.8 24.4 21 24.2 ...
##  $ pmRA : num  181.2 162.3 -44.2 367.1 157.7 ...
##  $ pmDE : num  -0.93 -62.4 -145.9 -19.49 -40.31 ...
##  $ e_Plx: num  3.1 0.78 1.97 1.71 1.36 1.28 1.91 1.22 1.64 2.17 ...
##  $ B.V  : num  0.999 0.778 1.15 1.03 1.068 ...
```

El dataset contiene 2719 datos con nueve conlumnas. Los nombres de las columnas son: HIP, RA, DE, Plx, pmRA, pmDE, e_Plx, B.V. Los datos que alberga son todos de tipo numérico.

- Muestra por pantalla la columna de la variable RA

```
head(hip$RA,n=30)
```

```
##  [1] 0.003797 0.111047 0.135192 0.151656 0.221873 0.243864 0.348708
##  [8] 0.426746 0.455182 0.478685 0.612287 0.696411 0.972063 1.099309
## [15] 1.102623 1.244275 1.281668 1.369764 1.423333 1.468617 1.843365
## [22] 1.966150 2.261459 2.315143 2.352249 2.431558 2.768701 2.878592
## [29] 2.898287 2.906145
```

- Calcula las tendencias centrales de todos los datos del dataset (mean, media) utilizando la function apply

```
apply(hip,2,mean)
```

```
##           HIP          Vmag            RA            DE           Plx
## 56549.4828981     8.2593858   173.4529975    -0.1397663    22.1980213
```

```
##          pmRA          pmDE         e_Plx           B.V
##     5.3761346   -63.9419934     1.6267929            NA
```

- Haz lo mismo para las medidas de dispersión mínimo y máximo. ¿Seria posible hacerlo con un único comando?¿Que hace la función range()

```
apply(hip,2,min,na.rm=TRUE)
```

```
##          HIP         Vmag           RA           DE          Plx
##     2.000000     0.450000     0.003797   -87.202730    20.000000
##         pmRA         pmDE        e_Plx          B.V
##  -868.010000 -1392.300000     0.450000    -0.158000
```

```
apply(hip,2,max,na.rm=TRUE)
```

```
##          HIP         Vmag           RA           DE          Plx
## 120003.00000     12.74000    359.95468     88.30268     25.00000
##         pmRA         pmDE        e_Plx          B.V
##    781.34000    481.19000     46.91000      2.80000
```

```
apply(hip,2,range,na.rm=TRUE)
```

```
##         HIP  Vmag         RA        DE Plx    pmRA     pmDE e_Plx    B.V
## [1,]      2  0.45   0.003797 -87.20273  20 -868.01 -1392.30  0.45 -0.158
## [2,] 120003 12.74 359.954685  88.30268  25  781.34   481.19 46.91  2.800
```

La función `range()` devuelve el máximo y el mínimo de los valores de cada columna.

- Sin embargo las medidas mas populares de dispersión son la varianza (var()), su desviación standard (sd()) y la desviación absoluta de la mediana o MAD. Calcula estas medidas para los valores de RA. Calculamos los valores de forma general.

```
apply(hip,2,var,na.rm=TRUE)
```

```
##          HIP         Vmag           RA           DE          Plx
## 1.266456e+09 3.552207e+00 1.156632e+04 1.515575e+03 2.008437e+00
##         pmRA         pmDE        e_Plx          B.V
## 2.591451e+04 1.985011e+04 4.896779e+00 1.012434e-01
```

```
apply(hip,2,sd,na.rm=TRUE)
```

```
##          HIP         Vmag           RA           DE          Plx
## 3.558731e+04 1.884730e+00 1.075468e+02 3.893039e+01 1.417193e+00
##         pmRA         pmDE        e_Plx          B.V
## 1.609799e+02 1.408904e+02 2.212867e+00 3.181876e-01
```

```
apply(hip,2,mad,na.rm=TRUE)
```

```
##          HIP         Vmag           RA           DE          Plx
## 4.909037e+04 1.882902e+00 1.469334e+02 4.398403e+01 1.764294e+00
##         pmRA         pmDE        e_Plx          B.V
## 1.416476e+02 9.949729e+01 4.892580e-01 2.809527e-01
```

Calcularmos los valores para la columna RA solamante.

```
var(hip$RA,na.rm = TRUE)
```

```
## [1] 11566.32
```

```
sd(hip$RA,na.rm = TRUE)
```

```
## [1] 107.5468
```

```
mad(hip$RA,na.rm = TRUE)
```

```
## [1] 146.9334
```

- Imagina que quieres calcular dos de estos valores de una sola vez. ¿Te serviría este código? f = function(x) c(median(x), mad(x))
  f(hip[,1])

```
f = function(x) c(median(x), mad(x))
f(hip[,1])
```

```
## [1] 56413.00 49090.37
```

Este código sí que nos serviría, a no ser que el la columna tuviera NAs, en ese caso devolvería NA. Para que esto no ocurra se debe cambiar lo siguiente dentro del código de la función.

```
f = function(x) c(median(x,na.rm=TRUE),mad(x,na.rm=TRUE))
```

- ¿Cuál sería el resultado de aplicar apply(hip,2,f)?

```
apply(hip,2,f)
```

```
##           HIP     Vmag       RA        DE       Plx     pmRA      pmDE
## [1,] 56413.00 8.280000 173.3698  3.254234 22.100000  10.5500 -49.48000
## [2,] 49090.37 1.882902 146.9334 43.984032  1.764294 141.6476  99.49729
##         e_Plx       B.V
## [1,] 1.140000 0.7105000
## [2,] 0.489258 0.2809527
```

- Vamos a medir la dispersión de la muestra utilizando el concepto de cuartiles. El percentil 90 es aquel dato que excede en un 10% a todos los demás datos. El cuartil (quantile) es el mismo concepto, solo que habla de proporciones en vez de porcentajes. De forma que el percentil 90 es lo mismo que el cuartil 0.90. La mediana "median" de un dataset es el valor más central, en otras palabras exactamente la mitad del dataset excede la media. Calcula el cuartil .10 y .50 para la columna RA del dataset hip. Sugerencia: quantile()

```
# Calculamos los cuartiles.
help("quantile")
quantile(hip$RA,probs=c(0.1,0.5))
```

```
##       10%       50%
##  28.92324 173.36979
```

- Los cuantiles 0.25 y 0.75 se conocen como el first quartile y el third quartile, respectivamente. Calcula los cuatro cuartiles para RA con un único comando.

```
quantile(hip$RA)
```

```
##         0%        25%        50%        75%       100%
##   0.003797  70.141368 173.369788 266.923319 359.954685
```

- Otra medida de dispersion es la diferencia entre el primer y el tercer cuartil conocida como rango intercuartil (IQR) Inter Quantile Range. ¿Obtienes ese valor con la función summary()?

```
summary(hip)
```

```
##       HIP             Vmag             RA                DE
##  Min.   :     2   Min.   : 0.450   Min.   :  0.0038   Min.   :-87.2027
##  1st Qu.: 21770   1st Qu.: 7.050   1st Qu.: 70.1414   1st Qu.:-31.3635
##  Median : 56413   Median : 8.280   Median :173.3698   Median :  3.2542
##  Mean   : 56549   Mean   : 8.259   Mean   :173.4530   Mean   : -0.1398
```

```
##  3rd Qu.: 87096   3rd Qu.: 9.610   3rd Qu.:266.9233   3rd Qu.: 28.0705
##  Max.   :120003   Max.   :12.740   Max.   :359.9547   Max.   : 88.3027
##
##       Plx              pmRA               pmDE              e_Plx
##  Min.   :20.00   Min.   :-868.010   Min.   :-1392.30   Min.   : 0.450
##  1st Qu.:20.98   1st Qu.: -91.980   1st Qu.: -130.79   1st Qu.: 0.870
##  Median :22.10   Median :  10.550   Median :  -49.48   Median : 1.140
##  Mean   :22.20   Mean   :   5.376   Mean   :  -63.94   Mean   : 1.627
##  3rd Qu.:23.36   3rd Qu.: 103.870   3rd Qu.:    8.57   3rd Qu.: 1.680
##  Max.   :25.00   Max.   : 781.340   Max.   :  481.19   Max.   :46.910
##
##       B.V
##  Min.   :-0.1580
##  1st Qu.: 0.5600
##  Median : 0.7105
##  Mean   : 0.7615
##  3rd Qu.: 0.9530
##  Max.   : 2.8000
##  NA's   :41
```

Con la función `summary()` no se obtiene el rango intercuartil. Si quisieramos obtener ese dato tendríamos que utilizar la función `IQR()`.

```
apply(hip,2,IQR,na.rm=TRUE)
```

```
##         HIP        Vmag          RA          DE         Plx        pmRA
## 65326.00000     2.56000   196.78195    59.43393     2.37500   195.85000
##        pmDE       e_Plx         B.V
##   139.36000     0.81000     0.39300
```

- Hasta ahora has ignorado la presencia de valores perdidos NA. La función any() devuelve TRUE si se encuentra al menos un TRUE en el vector que damos como argumento. Su combinación con is.na es muy útil. ¿qué obtienes cuando ejecutas el siguiente comando? ¿Cómo lo interpretas?
  hasNA = function(x) any(is.na(x))
  apply(hip,2,hasNA)

```
 hasNA = function(x) any(is.na(x))
 apply(hip,2,hasNA)
```

```
##   HIP  Vmag    RA    DE   Plx  pmRA  pmDE e_Plx   B.V
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

Solamente la columna "B.V" contiene missing values.

- Prueba a ejecutar el siguiente comando.
  hip1 = na.omit(hip)

```
hip1 = na.omit(hip)
```

- Como has observado nos devuelve NA para toda la columna, normalmente querríamos poder usar la función sobre el resto de datos que no son NA: Para ello podemos utilizar la función na.omit. ¿Que ocurre cuando lo hacemos?. Usando apply calcula la media para hip y hip1. Intenta calcular la media de forma que solo cambie la de B.V cuando ignores los valores NA.

```
apply(hip,2,mean)
```

```
##           HIP          Vmag            RA            DE           Plx
## 56549.4828981     8.2593858   173.4529975    -0.1397663    22.1980213
##          pmRA          pmDE         e_Plx           B.V
```

```
##     5.3761346   -63.9419934     1.6267929            NA
```

```
apply(hip1,2,mean)
```

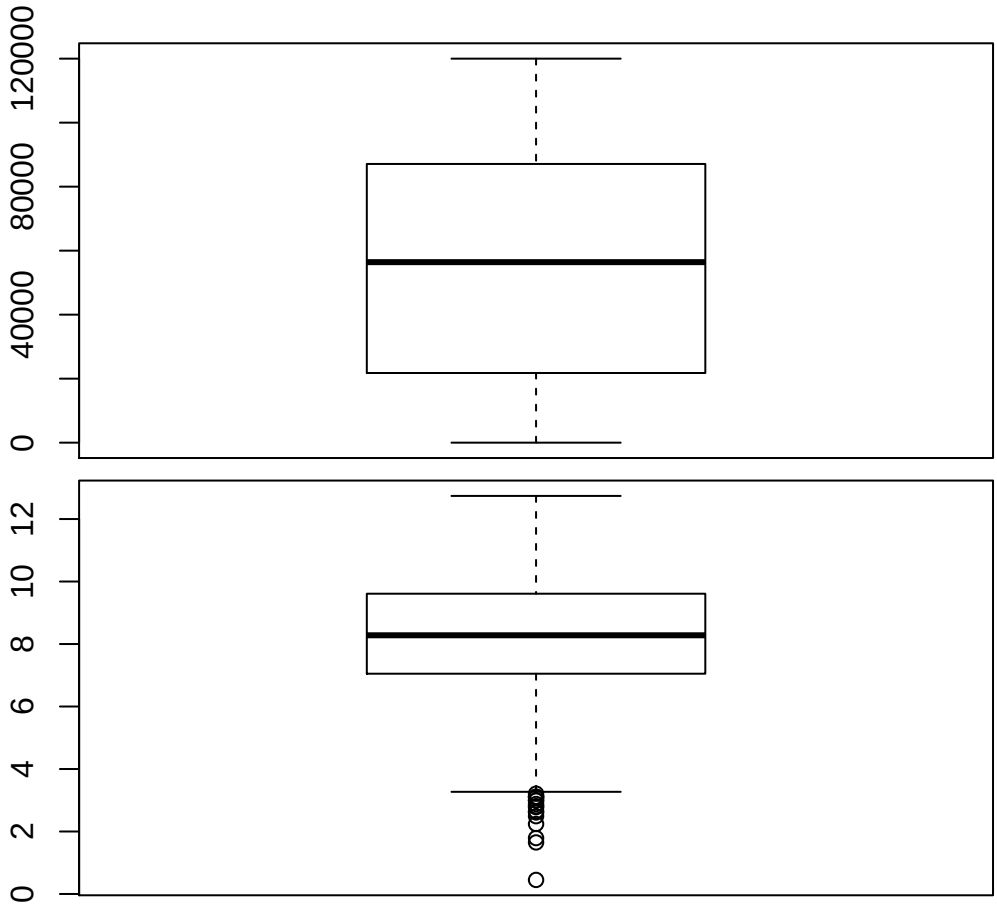
```
##           HIP          Vmag            RA            DE           Plx
## 56575.8050784     8.2147797   173.5284087    -0.2743560    22.1954033
##           pmRA          pmDE         e_Plx           B.V
##     5.5370575   -63.5345892     1.5449552     0.7615299
```

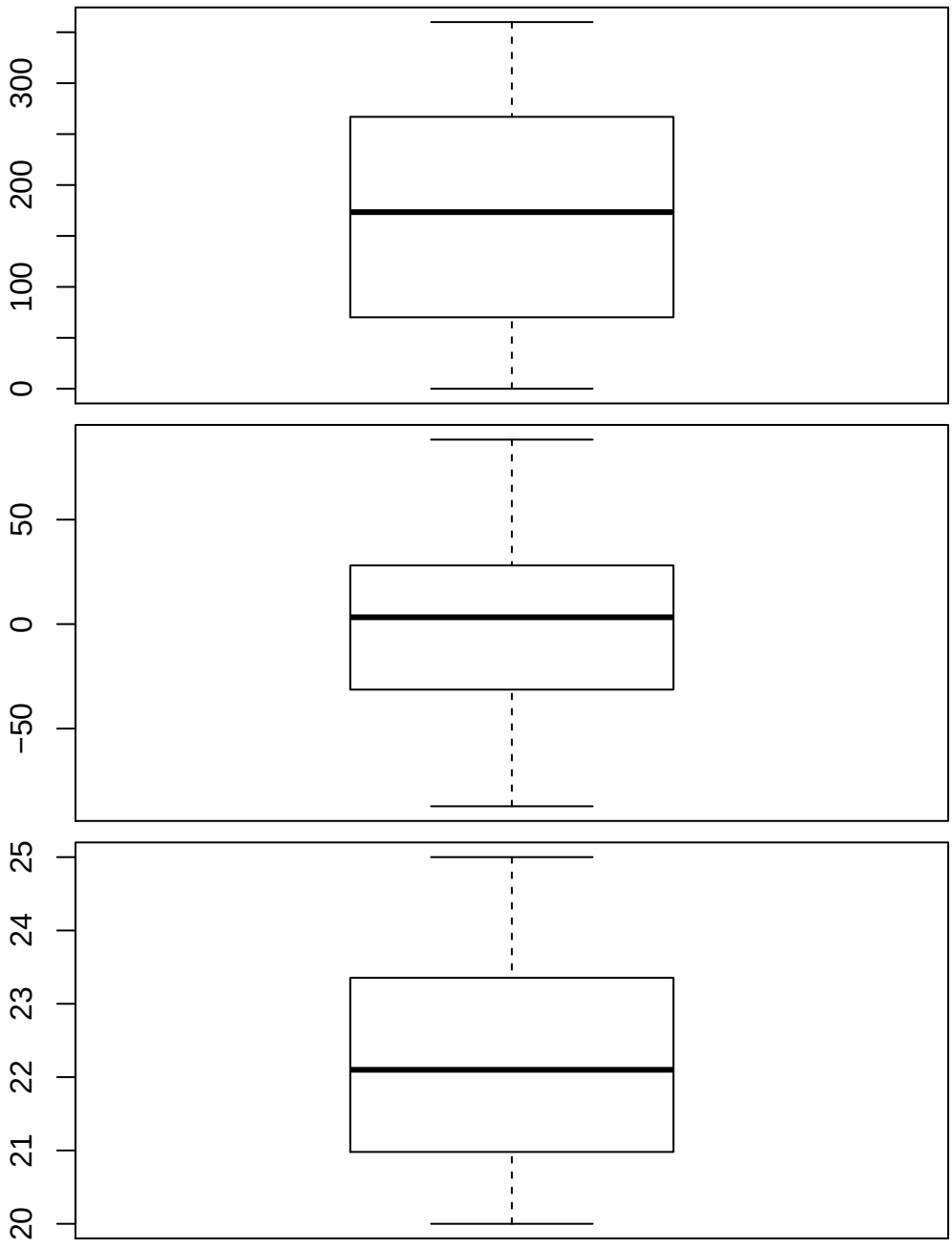
```
mean(hip$B.V,na.rm = TRUE)
```

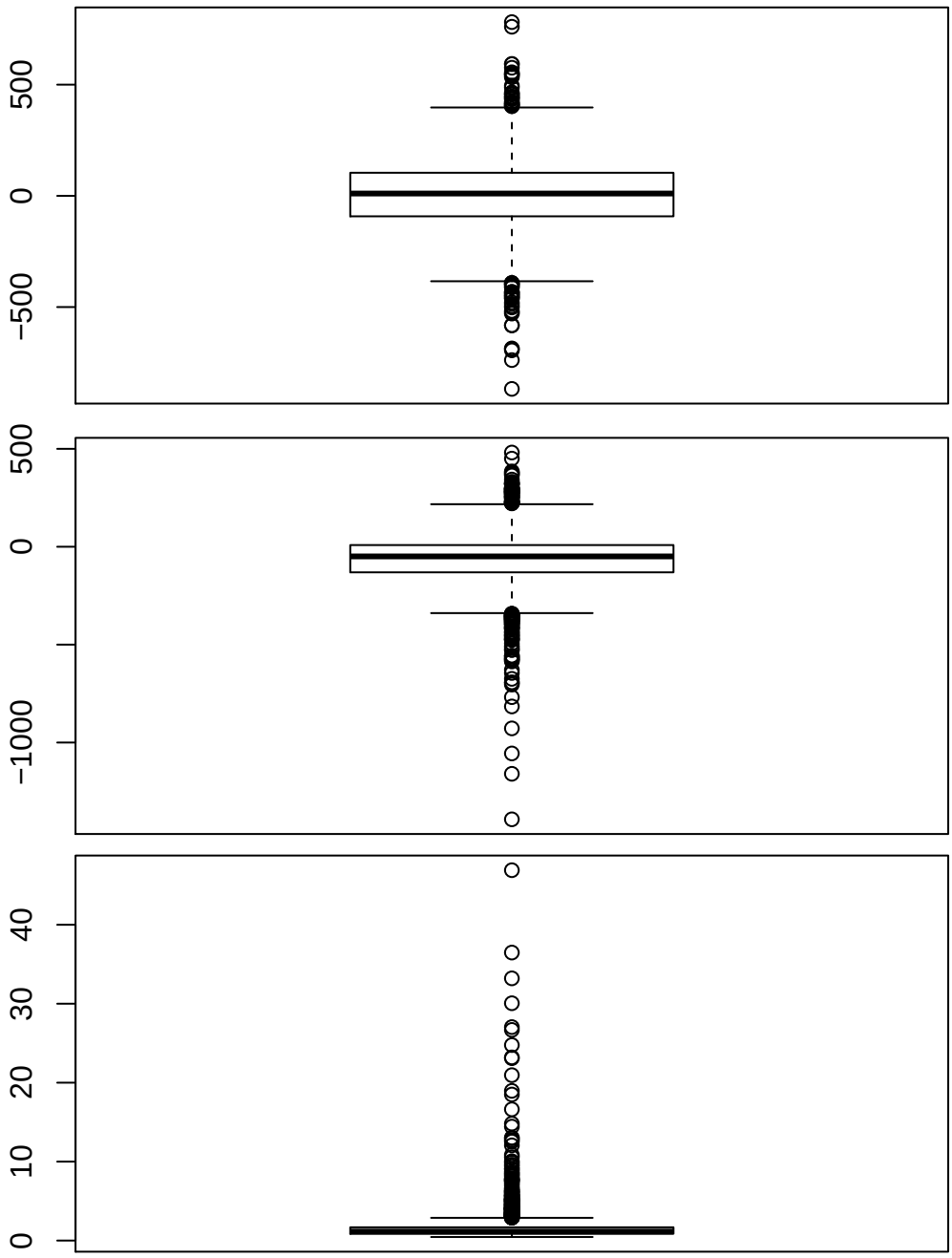
```
## [1] 0.7615299
```

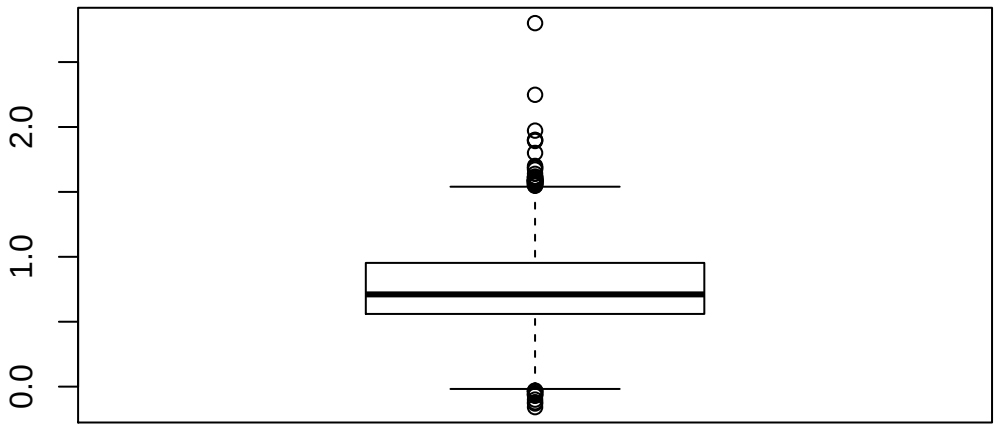
- Obten una idea aproximada de tus datos mediante la creación de un boxplot del hop dataset

```
colnames(hip)
apply(hip, 2, boxplot)
```




0
100
200
300
−50
0
50
20
21
22
23
24
25


500
0
−500
500
0
−1000
40
30
20
10
0



Gracias a los boxplot se puede ver que las columnas “Vmag”, “pmRA”, “pmDE”, “e_Plx” y “B.V” tienen outliers.

*Crea un scatterplot que te compare los valores de RA y DE. Representa los puntos con el símbolo ‘.’ Y que estos puntos sean de color rojo si DE excede de 0. Sugerencia ifelse()

- Haz un scatterplot de RA y pmRA. ¿Ves algún patrón?
- En vez de crear los plots por separado para cada par de columnas, hazlos con un solo comando con el scatterplot matrix
- Para poder acceder a las variables por su nombre usa attach(hip).Vamos a seleccionar las estrellas Hyadas del dataset aplicando los siguientes filtros:
  - RA in the range (50,100)
  - DE in the range (0,25)
  - pmRA in the range (90,130)
  - pmDE in the range (-60,-10)
  - e_Plx < 5
  - Vmag >4 OR B.V <0.2 (this eliminates 4 red giants)
- Crea un nuevo dataset con la aplicación de estos filtro. El Nuevo dataset se llama hyades. ¿Que dimensiones tiene? Grafica un scatterplot de Vmag vs B.V

b. Ejemplo 2, iris dataset

- Vamos a utilizar el ejemplo del dataset iris que está incluido en la distribución de R. Este dataset fue creado por Douglas Fisher. Consta de tres clases y tipos de 3 clases de tipos de flores:
  - *setosa*
  - *virginica*
  - *versicolor*

Cada una de ellas con cuatro atributos: + sepal width + sepal length + petal width + petal length

- Inspecciona las primeras filas del dataset y calcula el summary() del mismo con cada atributo del dataset
- Crea un histograma de petal.width , teniendo en cuenta que el numero de bins es variable fija este a 9. Añádele color y nombres al eje x “Petal Width”y al gráfico dale el nombre de “Histogram of Petal Width”. Crea un histograma para cada variable *Crea los cuartiles del dataset
- Representa en un boxplot la variable de ancho de hoja dependiendo del tipo de hoja que tengan
- Crea los cuartiles para cada tipo de iris y represéntalos en un plot como líneas cada una de un color
- Crea los boxplot de la longitud del pétalo en función de la especie de Iris.
- Compara con scatter plots las variables entre sí.

- El conjunto de datos "swiss" contiene una medida estandarizada de fecundidad y varios indicadores socioeconómicos para cada una de las 47 provincias francófonas de Suiza.

1. ¿Qué diagrama dibujaría para mostrar la distribución de todos los valores? ¿Qué conclusiones sacarías?
2. Dibuje gráficos para cada variable. ¿Qué puede concluir de las distribuciones con respecto a su forma y posibles valores atípicos?
3. Dibuja un diagrama de dispersión de Fertilidad frente a % Catholic. ¿Qué tipo de áreas tienen las tasas de fertilidad más bajas? 4.¿Qué tipo de relación existe entre las variables Educación y Agricultura?

- El conjunto de datos de aceites de oliva es bien conocido y se puede encontrar en varios paquetes, por ejemplo, como aceitunas en extracat.. La fuente original de los datos es el artículo [Forina et al., 1983].

1. Dibuje un scatterplot de las ocho variables continuas. ¿Cuáles de los ácidos grasos están fuertemente asociados positivamente y cuáles fuertemente asociados negativamente?
2. ¿Hay valores atípicos u otras características que valga la pena mencionar?

- El conjunto de datos se llama Lanza del paquete HSAUR2.

1. Se informan los datos de cuatro estudios. Dibuje un diagrama para mostrar si los cuatro estudios son igualmente grandes.
2. El resultado se mide por la clasificación de la variable con puntuaciones de 1 (mejor) a 5 (peor). ¿Cómo describirías la distribución?

- El paquete vcdExtra incluye datos de un viejo estudio de cáncer de mama sobre la supervivencia o muerte de 474 pacientes.

1. Convierta los datos en un data frame y dibuje gráficos para comparar las tasas de supervivencia, primero, por grado de malignidad y, en segundo lugar, por centro de diagnóstico.
2. ¿Qué diagrama dibujaría para comparar las tasas de supervivencia tanto por grado de malignidad como por centro de diagnóstico? ¿Importa el orden de las variables explicativas?

- Dataset Crabs (del paquete MASS) [Venables y Ripley, 2002]. Los autores inicialmente se transforman a una escala logarítmica y luego escriben que:

"The data are very highly correlated and scatterplot matrices and brush plots [i.e. interactive graphics] are none too revealing.".

Utilizando gráficos generales, comente si la transformación logaritmica fue una buena idea y si está de acuerdo con su afirmación sobre las correlaciones.