Análisis Exploratorio de Datos: Ejercicios

Antonio Manuel Milán Jiménez

10 de noviembre de 2018

Ejercicios:

```
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
```

Vemos que las dimensiones son 2719x9.

Estos son los nombres de las columnas:

```
colnames(hip)
```

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

Y aquí podemos observar las primeras columna del dataset, comprobando que todas las variables son numéricas:

```
hip[1:5,]
```

```
##   HIP  Vmag       RA        DE   Plx   pmRA    pmDE e_Plx   B.V
## 1   2  9.27 0.003797 -19.49884 21.90 181.21   -0.93  3.10 0.999
## 2  38  8.65 0.111047 -79.06183 23.84 162.30  -62.40  0.78 0.778
## 3  47 10.78 0.135192 -56.83525 24.45 -44.21 -145.90  1.97 1.150
## 4  54 10.57 0.151656  17.96896 20.97 367.14  -19.49  1.71 1.030
## 5  74  9.93 0.221873  35.75272 24.22 157.73  -40.31  1.36 1.068
```

Muestra por pantalla la columna de la variable RA

```
hip$RA[10]
```

```
## [1] 0.478685
```

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

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

```
##           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     0.7615299
```

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

```
##          HIP         Vmag           RA           DE          Plx
## 56413.000000     8.280000   173.369788     3.254234    22.100000
##         pmRA         pmDE        e_Plx          B.V
##    10.550000   -49.480000     1.140000     0.710500
```

```
apply(hip,2,mode)
```

```
##       HIP      Vmag        RA        DE       Plx      pmRA      pmDE
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##     e_Plx       B.V
## "numeric" "numeric"
```

Es interesante destacar que al ser datos numericos, no es posible calcular la moda sobre ellos, ya que no hay ninguno que se repita. También es necesario eliminar los "missing values" de la variable "B.V" para calcular los valores correctamente.

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,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
```

Podemos hacerlo en una sola línea ya que la función "range()" nos devovlerá un vector con el mínimo y el máximo de los datos proporcionados.