Lectura de datos

```
# Cargamos los tados
datos_train <- read.table("datos_train_preprocesado.csv", sep=",", comment.char="",quote = "\"", header=

datos_test <- read.table("datos_test_preprocesado.csv", sep=",", comment.char="",quote = "\"", header=TR
# Establecemos la semilla
```

Análisis exploratorio de los datos

El análisis exploratorio de datos o (EDA) engloba un conjunto de técnicas para poder comprender de manera rápida la naturaleza de una colección de datos o dataset.

Se basa principalmente en dos criterios: las **estadísticas de resumen** y la **visualización de datos**.

En primer lugar, vamos a realizar un resumen de nuestros datos utilizando la función `summary`. Dicha función nos mostrará información relevante para cada una de las columnas del datataset, mostrando información general como valores mínimos, máximos, media, mediana..

El resultado que obtenemos al evaluar nuestro dataset es el siguiente:

```
summary(datos_train)
```

```
##    urlDrugName       rating                      effectiveness
##  lexapro :  63   Min.   : 1.000   Considerably Effective: 926
##  prozac  :  46   1st Qu.: 5.000   Highly Effective      :1330
##  retin-a :  45   Median : 8.000   Ineffective           : 247
##  zoloft  :  45   Mean   : 7.008   Marginally Effective  : 186
##  paxil   :  38   3rd Qu.: 9.000   Moderately Effective  : 415
##  propecia:  38   Max.   :10.000
##  (Other) :2829
##                          sideEffects                 condition
##  Extremely Severe Side Effects: 175   depression         : 236
##  Mild Side Effects            :1019   acne               : 165
##  Moderate Side Effects        : 612   anxiety            :  63
##  No Side Effects              : 930   insomnia           :  54
##  Severe Side Effects          : 368   birth control      :  49
##                                       high blood pressure:  42
##                                       (Other)            :2495
##
##  none
##  None
##  NONE
##  None.
##  The treatment benefits were marginal at best.  Mood neither improved nor deteriorated, and anxiety w
##  Before the use of vagifem tablets, I had to endure a series of urinary infections after sometimes pa
##  (Other)
##                    sideEffectsReview sideEffectsNumber effectivenessNumber
##  none                         : 112    Min.   :1.000     Min.   :1.000
##  None                         :  73    1st Qu.:1.000     1st Qu.:3.000
##  None.                        :  19    Median :2.000     Median :4.000
##  No side effects.             :   9    Mean   :2.304     Mean   :3.936
##  There were no side effects.:   6    3rd Qu.:3.000     3rd Qu.:5.000
##  no side effects              :   5    Max.   :5.000     Max.   :5.000
```

```
##  (Other)                   :2880
##  weightedRating    ratingLabel     sideEffectsInverse
##  Min.   : 2.000   Min.   :0.0000   Min.   :1.000
##  1st Qu.: 6.000   1st Qu.:1.0000   1st Qu.:3.000
##  Median : 8.000   Median :1.0000   Median :4.000
##  Mean   : 7.528   Mean   :0.7874   Mean   :3.696
##  3rd Qu.:10.000   3rd Qu.:1.0000   3rd Qu.:5.000
##  Max.   :10.000   Max.   :1.0000   Max.   :5.000
##
##
##  none
##  lower blood pressur
##  prevent pregnanc
##  treatment benefit margin best mood neither improv deterior anxieti never signific allevi unsurpris
##  abl work without hassl im free pain right now there need cri anymor whenev urin
##  believ multipl treatment benefit take tylenol headach tylenol safe inexpens requir physician prescr
##  (Other)
##         effects_preprocesado
##  none              : 214
##  side effect       :  51
##  none notic        :  17
##  none awar         :   8
##  notic side effect:   7
##  none can tell     :   6
##  (Other)           :2801
```

A continuación se va a realizar un análisis de la información más relevante no textual, como el valor de **rating** de los usuarios, la **efectividad** y los **efectos secundarios** de dicho medicamento y por último, la **valoración ponderada del rating** teniendo en cuenta la proporción entre efectividad y efectos secundarios del medicamento.

Valoraciones de los medicamentos por parte de los usuarios.

En primer lugar vamos a analizar si el *rating* aportado por los usuarios sobre los medicamentos son buenos o no.

Empezamos obteniendo las frecuencias y porcentaje total de las valoraciones aportadas por los usuarios. Para ello se va a calcular la frecuencia de dicho atributo y su porcentaje respecto del total.

```
# Obtener frecuencias del rating
table(datos_train$rating)
```

```
##
##   1   2   3   4   5   6   7   8   9  10
## 305 102 146 107 158 157 349 558 480 742
```

```
# Calculamos el número de documentos
numDocuments <- dim(datos_train)[1]

# Calculamos el porcentaje de cada puntuación respecto del total.
table(datos_train$rating)/numDocuments
```

```
##
##          1          2          3          4          5          6
## 0.09826031 0.03286082 0.04703608 0.03447165 0.05090206 0.05057990
##          7          8          9         10
```

```
## 0.11243557 0.17976804 0.15463918 0.23904639
```

Como podemos observar, hay una mayoría de valoraciones positivas respecto a las negativas. De hecho el mayor porcentaje (casi el 24%) tienen la máxima valoración.

Podemos comprobar ésto mediante el uso de la moda.

```
# Función para calcular la moda. Se le pasa como parámetro un atributo
calcularModa<-function(var){
  frec.var<-table(var)
  valor<-which(frec.var==max(frec.var))  # Elementos con el valor m
  names(valor)
}

# Obtenemos la moda para el rating
calcularModa(datos_train$rating)
```

```
## [1] "10"
```

Como resumen en general del rating, se va a calcular la media y la mediana para calcular la tendencia central para dicha variable.

La media es la siguiente:

```
# Media
mean(datos_train$rating)
```

```
## [1] 7.008376
```

La mediana es la siguiente:

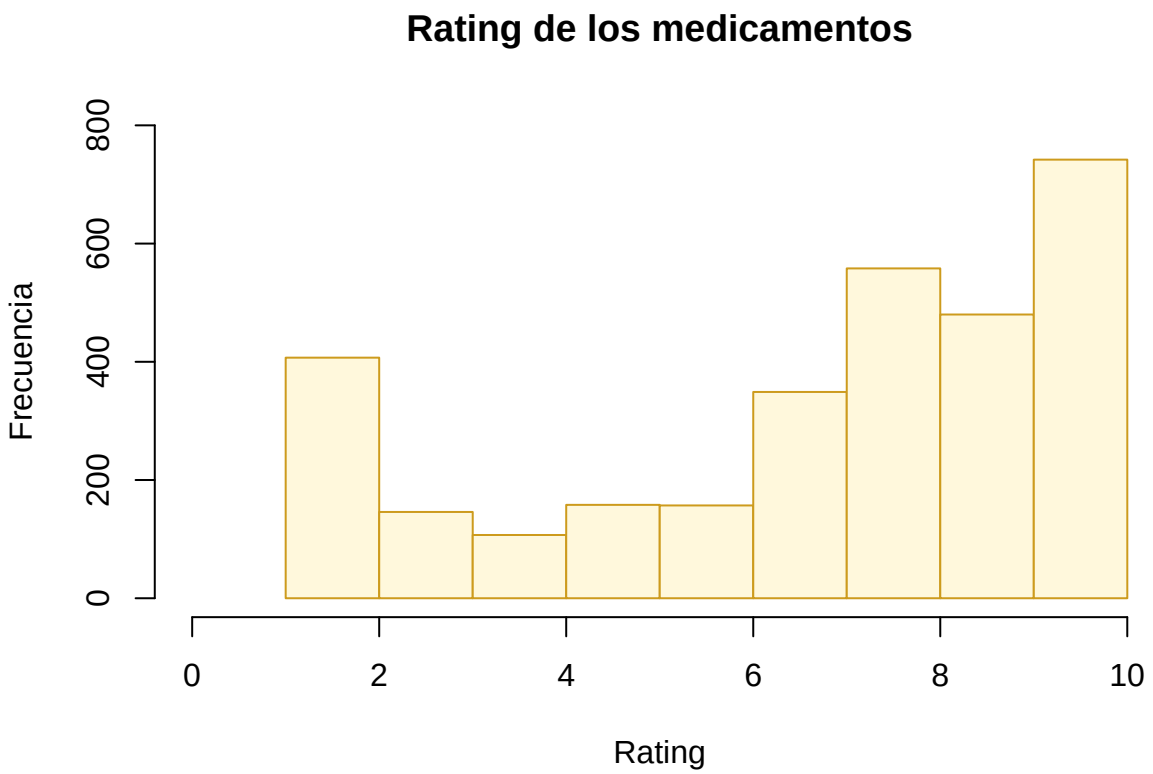
```
# Mediana
median(datos_train$rating)
```

```
## [1] 8
```

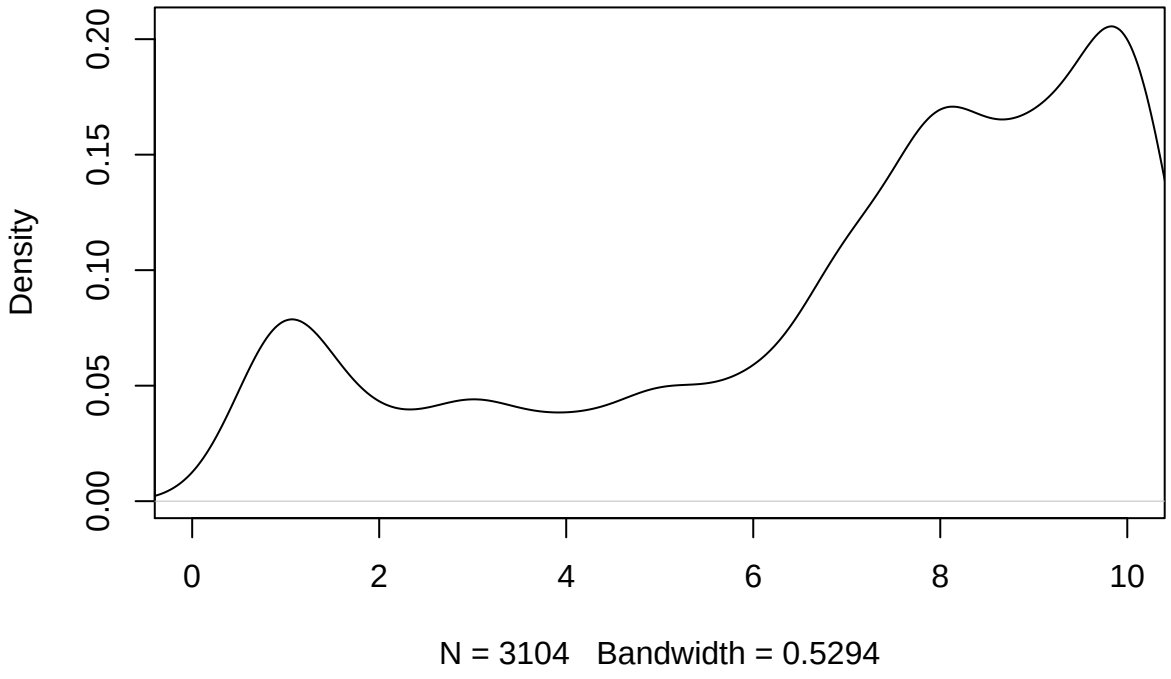
El valor obtenido es medio obtenido es 7 y la mediana es 8. Podemos concluir con dicha información, que en general las valoraciones sobre los medicamentos son bastante positivas, situándose el 50% de dichas valoraciones en el valor 8.

A continuación, se va a visualizar dicha información gráficamente:

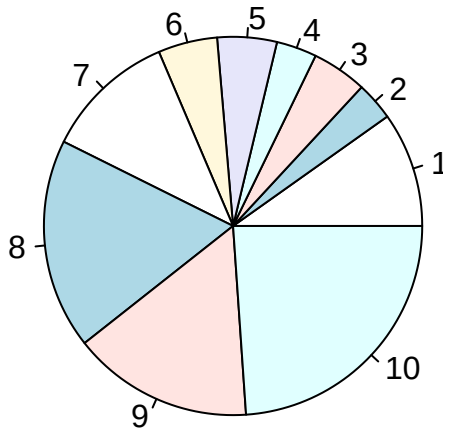
```
# Histograma de la valoración dada por los usuarios sobre los medicamentos

ratingExploration <- datos_train$rating

hist(ratingExploration,
     main="Rating de los medicamentos",
     xlab="Rating",
     ylab="Frecuencia",
     border="goldenrod3",
     xlim=c(0,10),
     ylim=c(0,800),
     col= "cornsilk",
     breaks=10,


    )
```



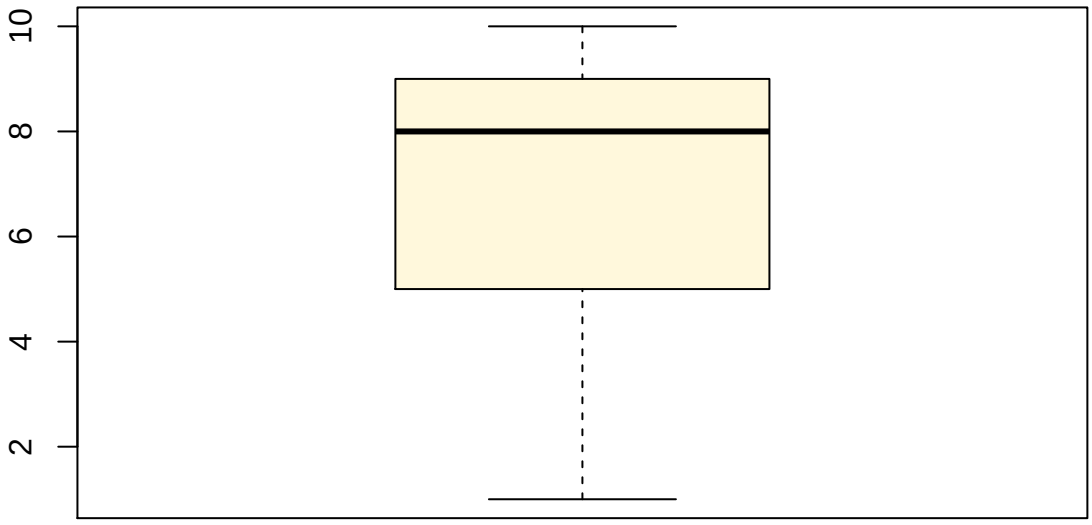

```
# Diagrama de densidad de la valoración dada por los usuarios sobre los medicamentos

plot(density(ratingExploration),
     main="Densidad del rating",
     xlim=c(0,10),
     )
```




```r
# Diagrama de sectores de las valoraciones dadas por los usuarios
pie(table(datos_train$rating))
```




```r
# Diagrama de cajas sobre las valoraciones dadas por los usuarios
boxplot(datos_train$rating,main="Rating", col= "cornsilk" )
```



Como medidas de dispersión, se va a calcular la **desviación típica**:

```
# Desviación típica
sd(datos_train$rating)
```

```
## [1] 2.937406
```

Como se puede observar, la desviación típica nos da un valor de 2.93. Esto quiere decir que los valores no están concentrados en un único valor, sino que la mayoría se sitúan en el un intervalo con distancia 3 despecto de la media.

Este valor concuerda, puesto que si observamos el histograma anterior, vemos que la mayoría de las puntuaciones se sitúan entre 5 y 10.

Esto también nos da como **conclusión** que en general las **opiniones** sobre los medicamentos **son buenas**, puesto que la mayor cantidad se sitúan en el intervalo [5.10].

Efectividad del medicamento

En esta sección, se va a analizar si se consideran que los medicamentos son efectivos o no. Para ello se va a analizar el atributo **effectivenessNumber** (que mide la efectividad del medicamento, siendo 1 menos efectivo y 5 más afectivo)

Empezamos obteniendo las frecuencias y porcentaje total de las anotaciones de efectividad. Para ello se va a calcular la frecuencia de dicho atributo y su porcentaje respecto del total.

```
# Obtener frecuencias del efectivenessNumber
table(datos_train$effectivenessNumber)
```

```
##
##    1    2    3    4    5
##  247  186  415  926 1330
```

```
# Calculamos el número de documentos
numDocuments <- dim(datos_train)[1]

# Calculamos el porcentaje de cada valor de efectividad respecto del total.
table(datos_train$effectivenessNumber)/numDocuments
```

```
##
##          1          2          3          4          5
## 0.07957474 0.05992268 0.13369845 0.29832474 0.42847938
```

Como podemos observar, la mayoría de los medicamentos se consideran que son efectivos. De hecho, la mayoría de los medicamentos se consideran altamente efectivos (con un 42%)

Podemos comprobar ésto mediante el uso de la moda.

```
# Obtenemos la moda para el efectivenessNumber
calcularModa(datos_train$effectivenessNumber)
```

```
## [1] "5"
```

Como resumen en general de la efectividad, se va a calcular la media y la mediana para calcular la tendencia central para dicha variable.

La media es la siguiente:

```
# Media
mean(datos_train$effectivenessNumber)
```

```
## [1] 3.936211
```

La mediana es la siguiente:

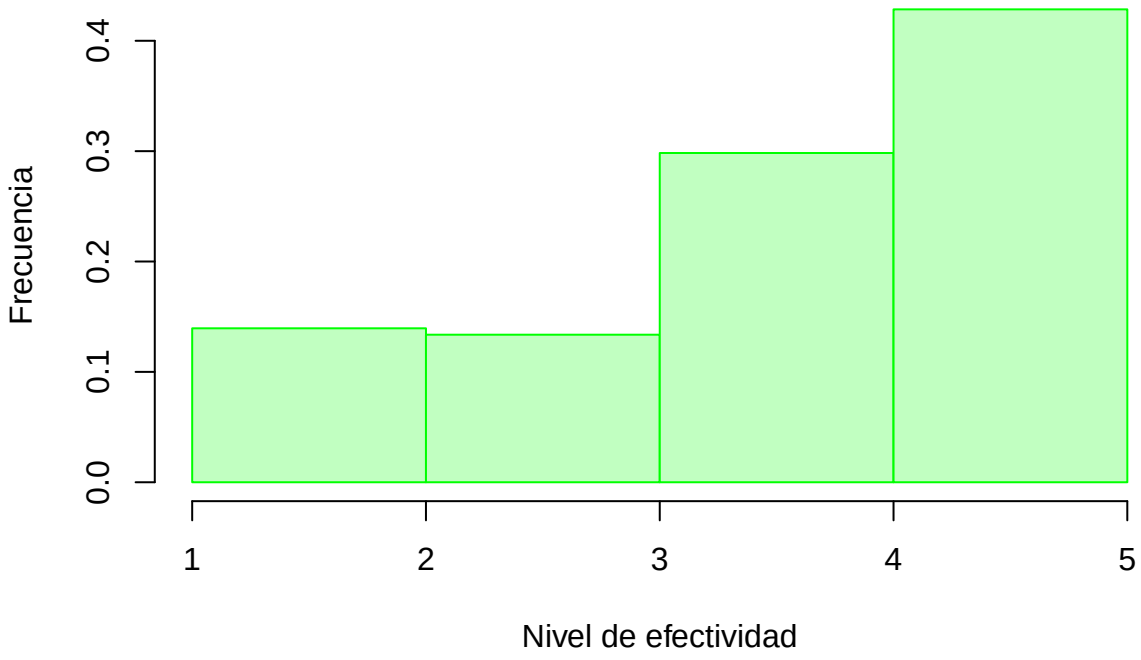
```
# Mediana
median(datos_train$effectivenessNumber)
```

```
## [1] 4
```

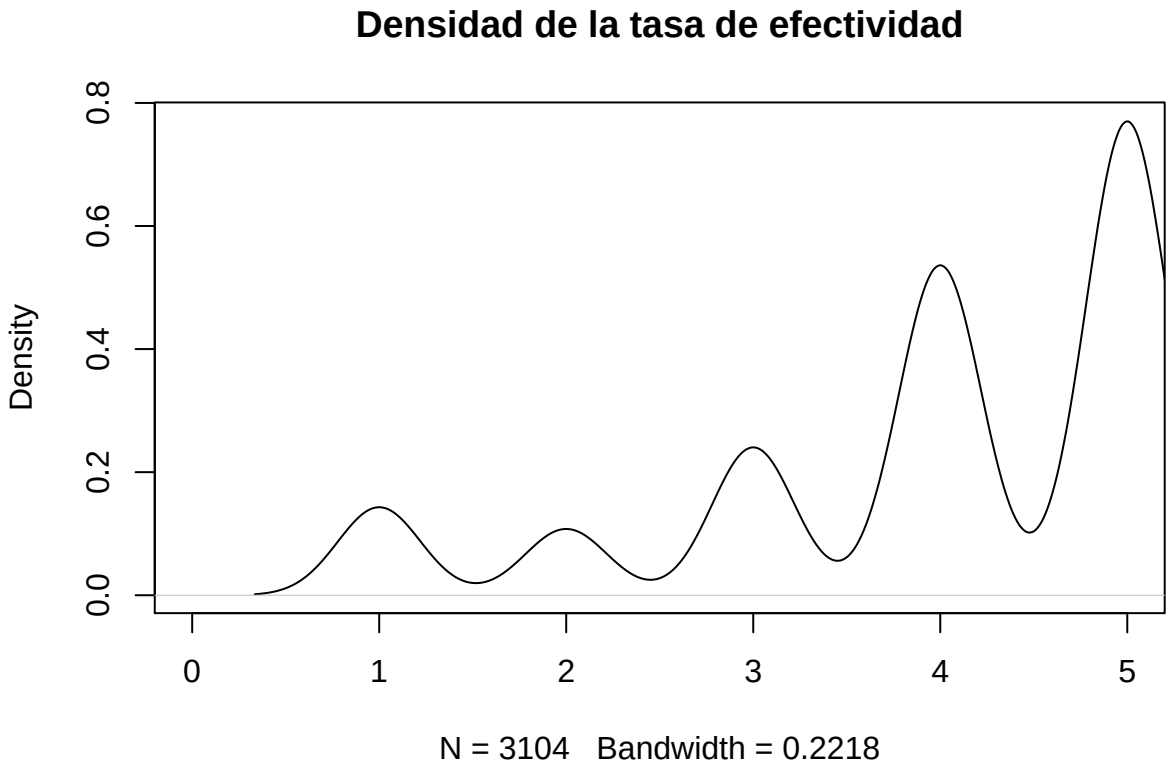
El valor medio obtenido es 3.93 sobre 5 y la mediana es 4. Podemos concluir con dicha información, que en general los medicamentos son bastantes efectivos, situándose el 50% de dichas mediciones sobre el valor 4.

A continuación, se va a visualizar dicha información gráficamente:

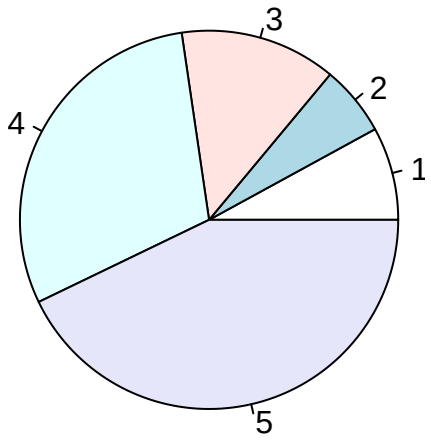
```
# Histograma sobre la efectividad de los medicamentos

efecctivenessNumberExploration <- datos_train$effectivenessNumber

hist(efecctivenessNumberExploration,
     main="Efectividad de los medicamentos",
     xlab="Nivel de efectividad",
     ylab="Frecuencia",
     border="green",
     xlim=c(1,5),
     col= "darkseagreen1",
     breaks=5,
     prob=TRUE
    )
```



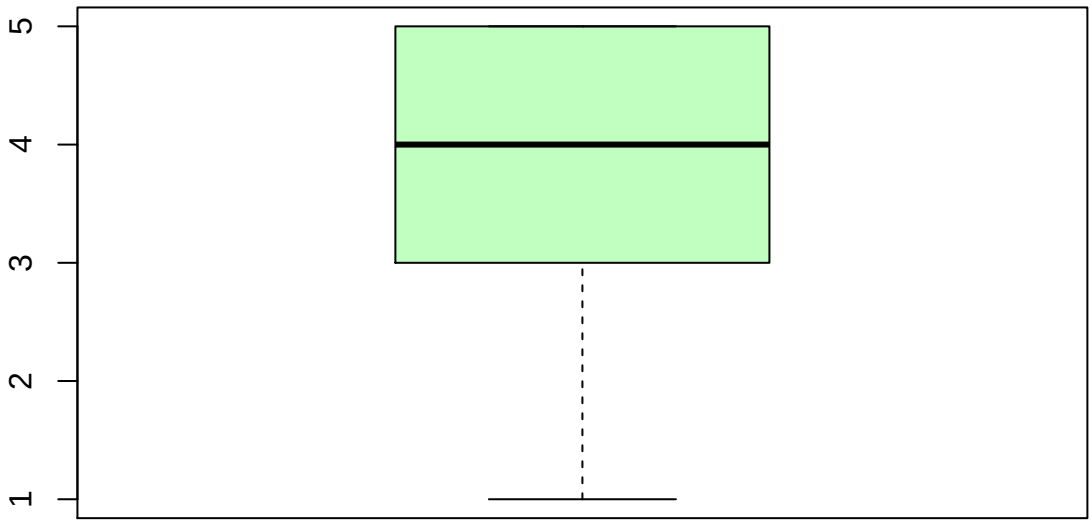

```
# Diagrama de densidad de la efectividad de los medicamentos

plot(density(datos_train$effectivenessNumber),
     main="Densidad de la tasa de efectividad",
     xlim=c(0,5),
     )
```




```r
# Diagrama de sectores de la efectividad de los medicamentos
pie(table(datos_train$effectivenessNumber))
```




```r
# Diagrama de cajas sobre la efectividad de los medicamentos
boxplot(datos_train$effectivenessNumber,main="Tasa de efectividad", col= "darkseagreen1" )
```



Como medidas de dispersión, se va a calcular la **desviación típica**:

```
# Desviación típica
sd(datos_train$effectivenessNumber)
```

```
## [1] 1.230634
```

Como se puede observar, la desviación típica nos da un valor de 1.23. Esto quiere decir que la mayor parte de los valores se sitúan en un intervalo con una distancia de uno de la media.

Este valor concuerda, puesto que si observamos el histograma anterior, vemos que la mayoría de las puntuaciones se sitúan entre 3 y 5.

Esto también nos da como **conclusión** que en general los medicamentos tienen una tasa bastante **buena de efectividad** puesto que su tasa se sitúa entre [3.5].

Efectos secundarios del medicamento

En esta sección, se va a analizar si se consideran que los medicamentos tienen efectos secundarios o no. Para ello se va a analizar el atributo **sideEffectsNumber** (que mide la tasa de efectos secundarios del medicamento, siendo 1 el mínimo de efectos secundarios y 5 el máximo de efectos secundarios)

Empezamos obteniendo las frecuencias y porcentaje total de las anotaciones de efectos secundarios. Para ello se va a calcular la frecuencia de dicho atributo y su porcentaje respecto del total.

```
# Obtener frecuencias del sideEffectsNumber
table(datos_train$sideEffectsNumber)
```

```
##
##    1    2    3    4    5
##  930 1019  612  368  175
```

```
# Calculamos el número de documentos
numDocuments <- dim(datos_train)[1]

# Calculamos el porcentaje de tasa de efectos secundarios respecto del total.
table(datos_train$sideEffectsNumber)/numDocuments
```

```
##
##          1          2          3          4          5
## 0.29961340 0.32828608 0.19716495 0.11855670 0.05637887
```

Como podemos observar, la mayoría de los medicamentos se consideran que no tienen efectos secundarios severos. De hecho, la mayoría de los medicamentos se sitúan entre sin efectos secundarios (29%) o que tienen efectos secundarios leves (32%).

Podemos comprobar ésto mediante el uso de la moda.

```
# Obtenemos la moda para el sideEffectsNumber
calcularModa(datos_train$sideEffectsNumber)
```

```
## [1] "2"
```

Como resumen en general de sobre la tasa de efectos secundarios, se va a calcular la media y la mediana para calcular la tendencia central para dicha variable.

La media es la siguiente:

```
# Media
mean(datos_train$sideEffectsNumber)
```

```
## [1] 2.303802
```

La mediana es la siguiente:

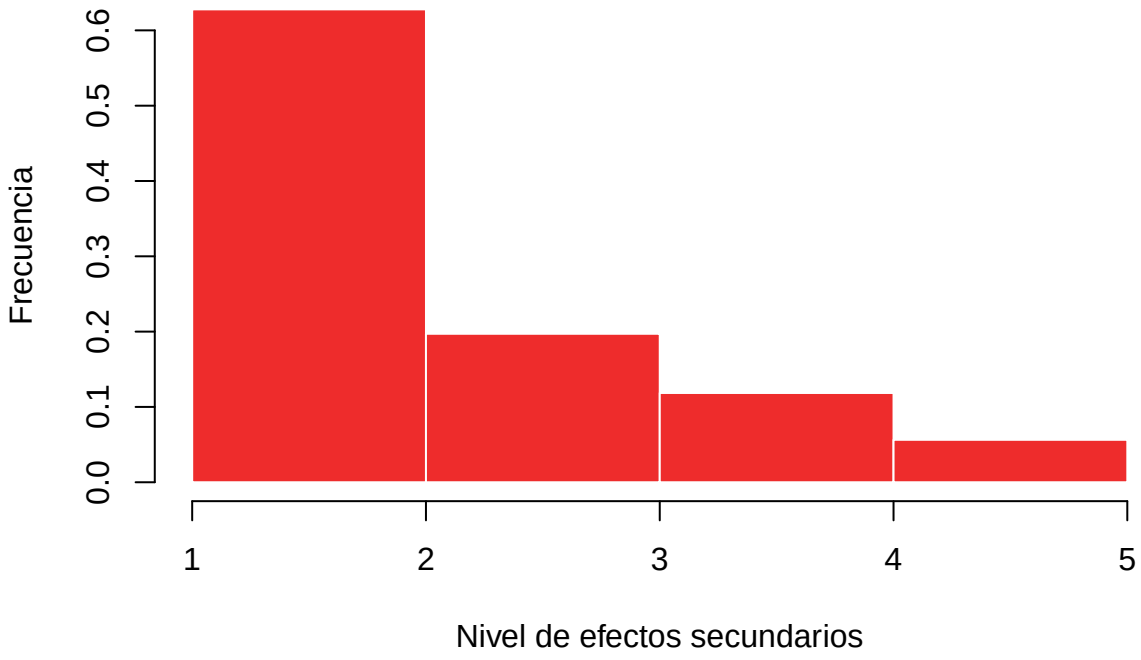
```
# Mediana
median(datos_train$sideEffectsNumber)
```

```
## [1] 2
```

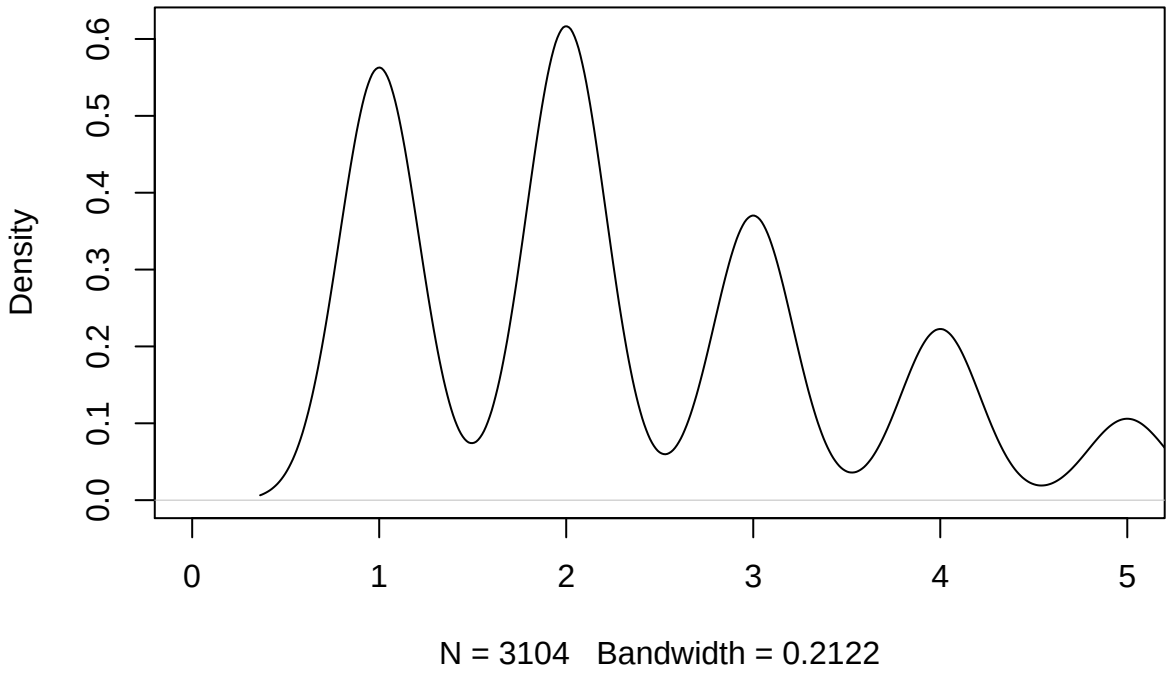
El valor medio obtenido es 2.30 sobre 5 y la mediana es 2. Podemos concluir con dicha información, que en general los medicamentos no tienen efectos secundarios o que dichos efectos son leves.

A continuación, se va a visualizar dicha información gráficamente:

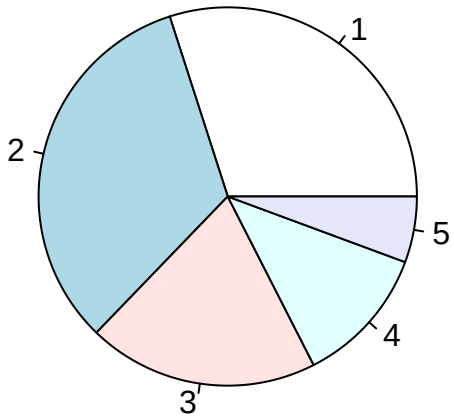
```
# Histograma de la tasa de efectos secundarios

sideEffectsNumberExploration <- datos_train$sideEffectsNumber

hist(sideEffectsNumberExploration,
     main="Efectos secundarios de los medicamentos",
     xlab="Nivel de efectos secundarios",
     ylab="Frecuencia",
     border="white",
     xlim=c(1,5),
     col= "firebrick2",
     breaks=5,
     prob=TRUE
    )
```



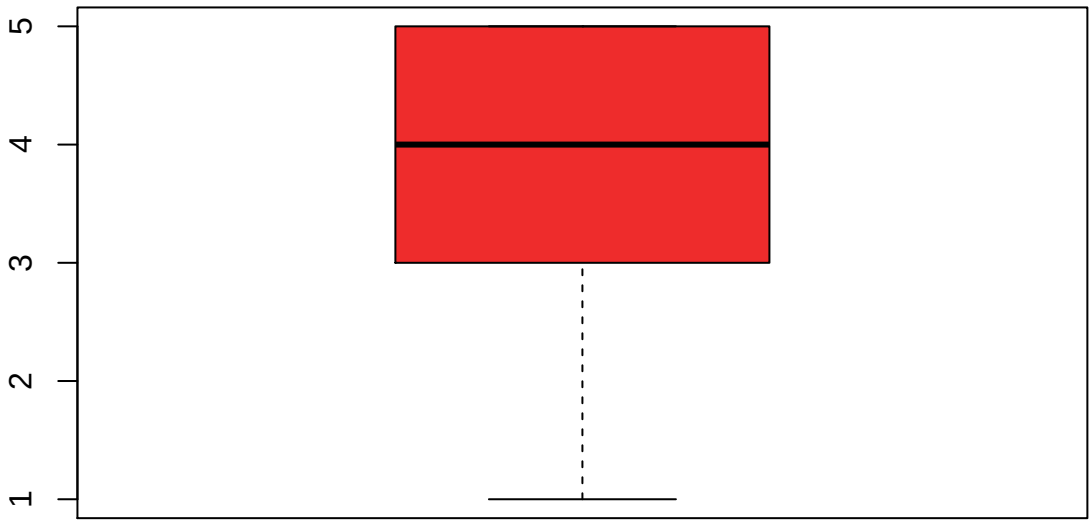

```r
# Diagrama de densidad sobre la tasa de efectos secundarios

plot(density(datos_train$sideEffectsNumber),
     main="Densidad de la tasa de efectos secundarios",
     xlim=c(0,5),
     )
```




```r
# Diagrama de sectores de los efectos secundarios de los medicamentos
pie(table(datos_train$sideEffectsNumber))
```




```r
# Diagrama de cajas de los efectos secundarios de los medicamentos
boxplot(datos_train$effectivenessNumber,main="Tasa de efectos secundarios", col= "firebrick2" )
```



Como medidas de dispersión, se va a calcular la **desviación típica**:

```
# Desviación típica
sd(datos_train$sideEffectsNumber)
```

```
## [1] 1.177525
```

Como se puede observar, la desviación típica nos da un valor de 1.17. Esto quiere decir que la mayor parte de los valores se sitúan en un intervalo con una distancia de uno de la media.

Este valor concuerda, puesto que si observamos el histograma anterior, vemos que la mayoría de las puntuaciones se sitúan entre 1 y 2.

Esto también nos da como **conclusión** que en general los medicamentos **no tienen efectos secundarios o son muy leves**.

Valoración ponderada sobre el medicamento

En esta sección, se va a analizar si se consideran que los medicamentos son buenos o no teniendo en cuenta la relación entre los beneficios que aporta (efectividad) y las inconvenientes que tiene (efectos secundarios). Para ello se va a analizar el atributo **weightedRating** (que mide dicha relación teniendo en cuenta una tasa de efectividad del 30% y una tasa de efectos secundarios del 70%), y siendo 1 peor valorado y 10 mejor valorado.

Empezamos obteniendo las frecuencias y porcentaje total de las anotaciones sobre la puntuación ponderada. Para ello se va a calcular la frecuencia de dicho atributo y su porcentaje respecto del total.

```
# Obtener frecuencias del weightedRating
table(datos_train$weightedRating)
```

```
##
##    2    4    6    8   10
##   92  327  587 1314  784
```

```
# Calculamos el número de documentos
numDocuments <- dim(datos_train)[1]

# Calculamos el porcentaje de puntuación ponderada respecto del total.
table(datos_train$weightedRating)/numDocuments
```

```
##
##          2          4          6          8         10
## 0.02963918 0.10534794 0.18911082 0.42332474 0.25257732
```

Como podemos observar, la mayoría de los medicamentos se consideran que son generalmente beneficiosos. De hecho, la mayoría de los medicamentos se sitúan con una valoración de 8 sobre 10 teniendo en cuenta la relación beneficio/perjuicio.

Podemos comprobar ésto mediante el uso de la moda.

```
# Obtenemos la moda para el weightedRating
calcularModa(datos_train$weightedRating)
```

```
## [1] "8"
```

Como resumen en general de sobre la tasa de efectos secundarios, se va a calcular la media y la mediana para calcular la tendencia central para dicha variable.

La media es la siguiente:

```
# Media
mean(datos_train$weightedRating)
```

```
## [1] 7.527706
```

La mediana es la siguiente:

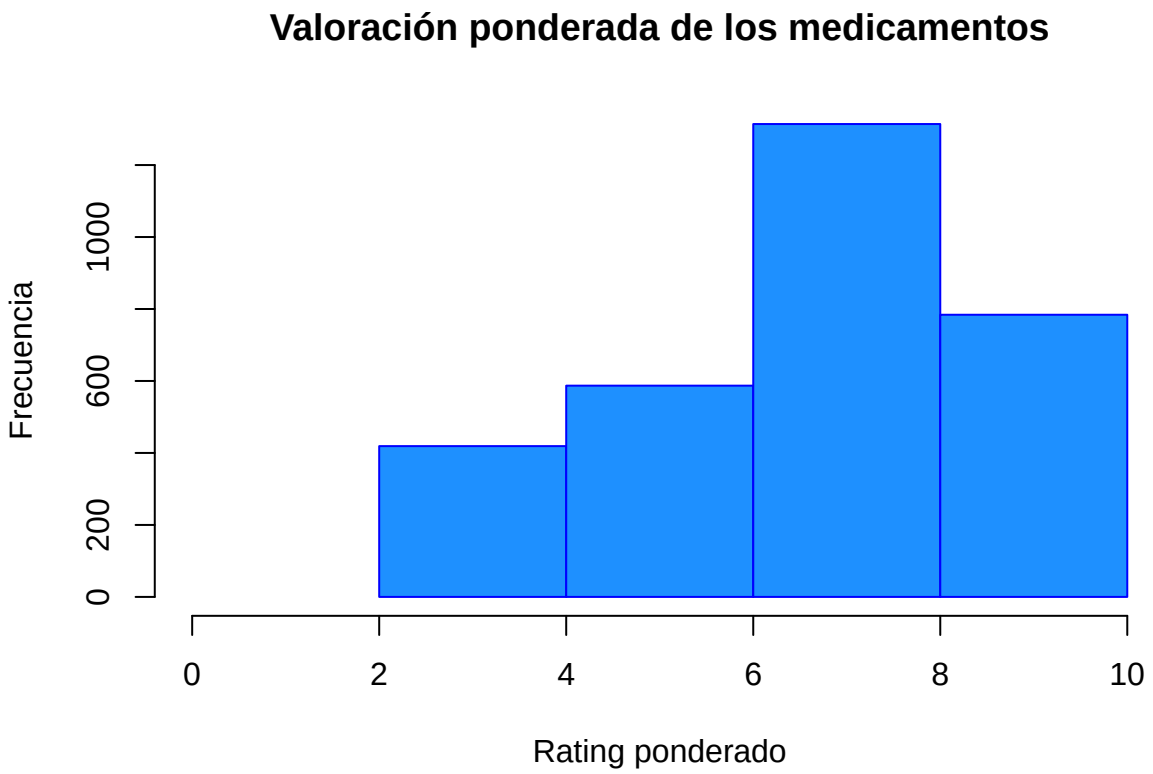
```
# Mediana
median(datos_train$weightedRating)
```

```
## [1] 8
```

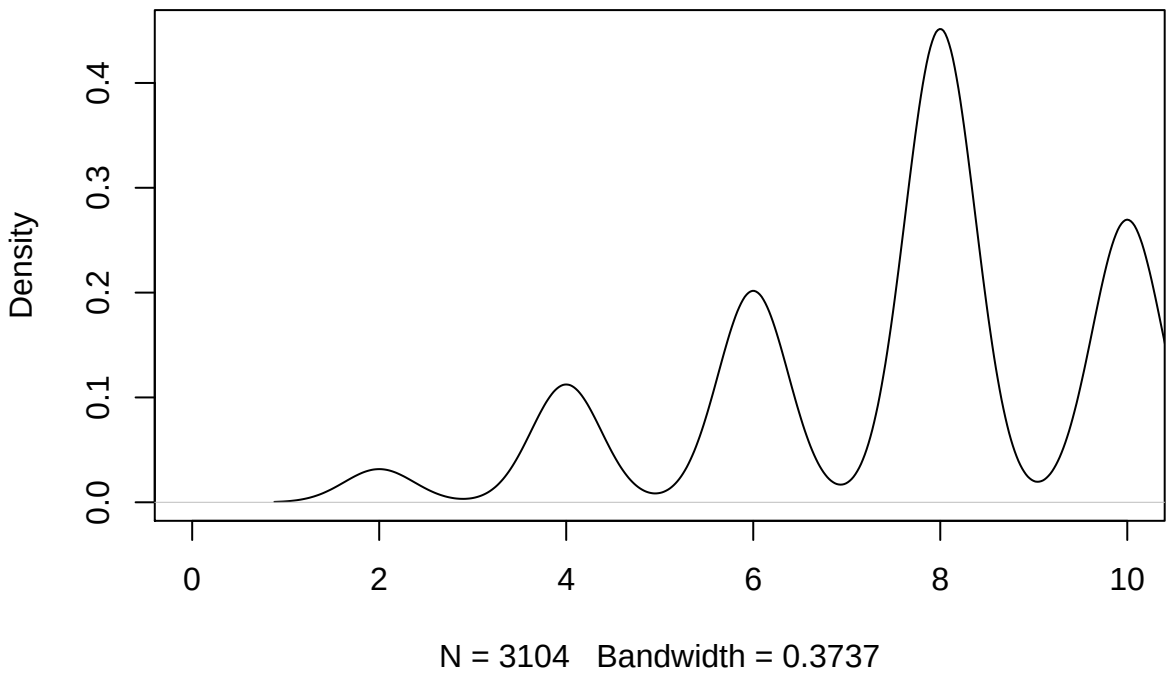
El valor medio obtenido es 7.52 sobre 10 y la mediana es 2. Podemos concluir con dicha información que la puntuación general sobre los medicamentos es de **notable**.

A continuación, se va a visualizar dicha información gráficamente:

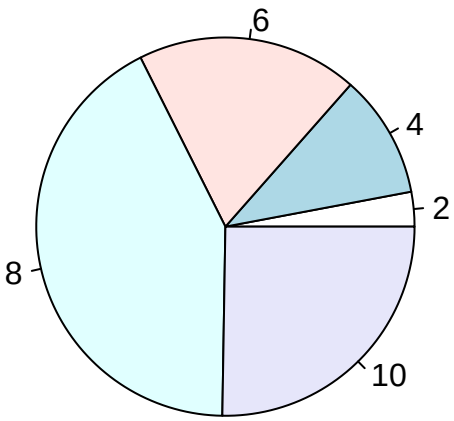
```
# Histograma de valoración ponderada

weightedRatingExploration <- datos_train$weightedRating

hist(weightedRatingExploration ,
     main="Valoración ponderada de los medicamentos",
     xlab="Rating ponderado",
     ylab="Frecuencia",
     border="blue",
     xlim=c(0,10),
     col= "dodgerblue1",
     breaks=5

    )
```



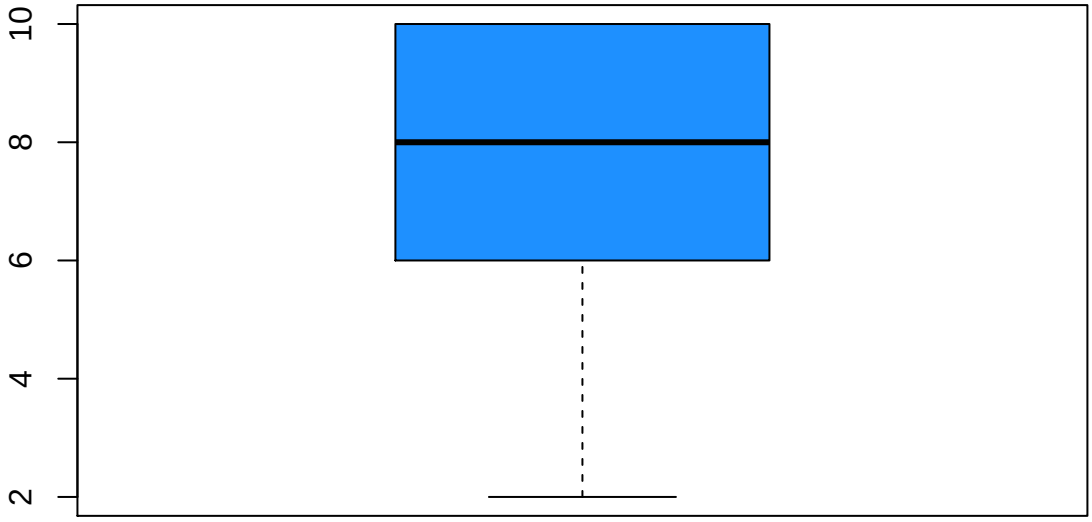

```r
# Diagrama de densidad sobre la valoración ponderada

plot(density(datos_train$weightedRating),
     main="Densidad de valoración ponderada",
     xlim=c(0,10),
     )
```




```r
# Diagrama de sectores sobre la valoración ponderada
pie(table(datos_train$weightedRating))
```




```r
# Diagrama de cajas sobre la valoración ponderada
boxplot(datos_train$weightedRating,main="Valoración ponderada", col= "dodgerblue1" )
```



Como medidas de dispersión, se va a calcular la **desviación típica**:

```
# Desviación típica
sd(datos_train$weightedRating)
```

```
## [1] 2.073078
```

Como se puede observar, la desviación típica nos da un valor de 2.07. Esto quiere decir que la mayor parte de los valores se sitúan en un intervalo con una distancia de dos de la media.

Este valor concuerda, puesto que si observamos el histograma anterior, vemos que la mayoría de las puntuaciones se sitúan entre 6 y 10.

Esto también nos da como **conclusión** que en general los medicamentos **son convenientes tomarlos**.

Correlación sobre las variables

En esta sección se va a comprobar la correlación que existe entre las variables que miden la efectividad(effectivenessNumber), los efectos secundarios(sideEffectsNumber), la valoración aportada por los usuarios(rating) y la valoración ponderada que se ha realizado sobre el medicamento(weightedRating).

Empezamos calculando la correlación entre la variable que mide la efectividad y los efectos secundarios.

```
# Correlación lineal entre efectivenessNumber y sideEffectsNumber
cor(datos_train[,c(8,9)])
```

```
##                     sideEffectsNumber effectivenessNumber
## sideEffectsNumber           1.0000000          -0.3953789
## effectivenessNumber        -0.3953789           1.0000000
```

Podemos observar que cuantos más efectos secundarios tiene, menor es la efectividad del medicamento. Esto puede estar influido por las valoraciones subjetivas del usuario, ya que si ha tenido una mala experiencia (debido a los efectos secundarios) por la ingesta del medicamento, no va a hacer énfasis en los beneficios del medicamento, sino que hará un mayor énfasis en los aspectos negativos.

A continuación vamos a calcular la correlación entre la efectividad y la valoración ponderada del medicamento.

```r
# Correlación lineal entre efectivenessNumber y weightedRating
cor(datos_train[,9:10])
```

```
##                      effectivenessNumber weightedRating
## effectivenessNumber           1.0000000      0.6561697
## weightedRating                0.6561697      1.0000000
```

Como se puede observar, cuando el medicamento es más efectivo, la valoración ponderada del medicamento acumenta (como es obvio), y si ahora calculamos la valoración ponderada del medicamento teniendo en cuenta los efectos secundarios

```r
# Correlación lineal entre sideEffectsNumber y weightedRating
cor(datos_train[,c(8,10)])
```

```
##                    sideEffectsNumber weightedRating
## sideEffectsNumber          1.0000000     -0.9165498
## weightedRating            -0.9165498      1.0000000
```

Observamos como si el medicamento tiene una mayor tasa de efectos secundarios, la valoración ponderada disminuye considerablemente (obvio porque el 70% de la valoración ponderada tiene en cuenta los efectos secundarios del medicamento).

Ahora vamos a comprobar la relación que existe entre la valoración dada por el usuario y los efectos secundarios.

```r
# Correlación lineal entre rating y sideEffectsNumber
cor(datos_train[,c(2,8)])
```

```
##                       rating sideEffectsNumber
## rating             1.000000         -0.682939
## sideEffectsNumber -0.682939          1.000000
```

Podemos comprobar como si el medicamento tiene una mayor tasa de efectos secundarios, la valoración dada por el usuario disminuye

```r
# Correlación lineal entre rating y effectivenessNumber
cor(datos_train[,c(2,9)])
```

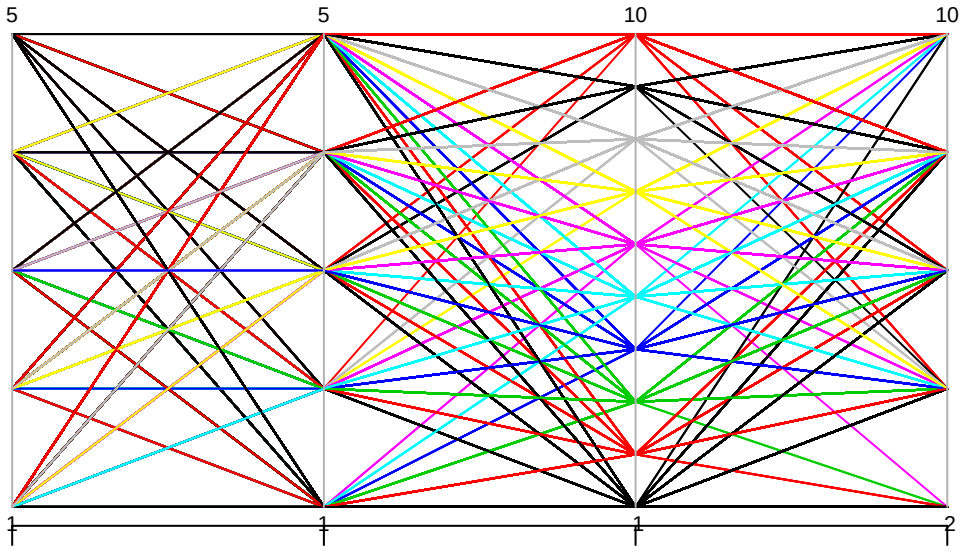
```
##                         rating effectivenessNumber
## rating              1.0000000           0.7498171
## effectivenessNumber 0.7498171           1.0000000
```

Y si la efectividad del medicamento es alta, la valoración del usuario se incrementa.

Como observación general, se puede destacar que **la valoración del usuario está condicionada más por la efectividad del medicamento** (relación 1/0.74) que por los efectos secundarios (relación 1/-0.68).

Por último, vamos a observar en el siguiente gráfico como se relacionan las variables entre sí en función se sus valores.

```r
# Gráfico de coordenadas paralelas
library(MASS)
parcoord(datos_train[,c(8,9,2,10)], col=datos_train$rating,var.label=T)
```



Por ejemplo, podemos destacar como si el número de efectos secundarios es 1 (no tiene efectos secundarios) y la efectividad del medicamento es 5 (muy efectivo), entonces la valoración del usuario será 10 y la valoración ponderada será también 10.