Análisis Exploratorio de Datos con R

Gaston Nina Sossa

Contents

Cargando el Dataset . . . 1
Frecuencia y Moda . . . 2
Medidas de Tendencia Central . . . 2

Comparación entre la moda, la mediana y la media **3**
Percentiles . . . 3
Resumen del Data Frame . . . 3
Análisis por especie . . . 4
Medidas de Dispersión . . . 5
Estadísticas Multivariadas . . . 5
Tablas de Contingencia . . . 6
Visualización . . . 6
Densidad . . . 10
Pie Chart . . . 11
Boxplots . . . 12
Diagramas de Disperción . . . 16
Gráficos de Coordenadas Paralelas . . . 22

El Dataset IRIS

El dataset se compone de 150 observaciones de flores de la planta iris.

Cargando el Dataset

```
data(iris)
names(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Acceder a las variables

```
attach(iris)
```

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Frecuencia y Moda

```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

```
vec <- c(1,1,1,0,0,3,3,3,3,2)
```

```
table(vec)
```

```
## vec
## 0 1 2 3
## 2 3 1 4
```

```
table(vec)/length(vec)
```

```
## vec
##   0   1   2   3
## 0.2 0.3 0.1 0.4
```

```
my_mode <- function(var){
  frec.var <- table(var)
  valor <- which(frec.var == max(frec.var))
  names(valor)
}
```

```
my_mode(vec)
```

```
## [1] "3"
```

```
my_mode(iris$Sepal.Length)
```

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

Medidas de Tendencia Central

```
vec <- rnorm(10,20,10)
vec
```

```
##  [1] 28.6313806  2.6029274 16.6884214 26.6121197  0.9389393 32.2137608
##  [7]  9.2538300 13.8347110 26.5872548 25.6307701
```

```
mean(vec)
```

```
## [1] 18.29941
```

```
vec.ruid <- c(vec, rnorm(1,300,100))
vec.ruid
```

```
##  [1]  28.6313806   2.6029274  16.6884214  26.6121197   0.9389393  32.2137608
##  [7]   9.2538300  13.8347110  26.5872548  25.6307701 395.9127460
```

```
mean(vec.ruid)
```

```
## [1] 52.6279
```

```
mean(vec, trim = 0.1)
```

```
## [1] 18.73018
```

```r
mean(vec.ruid, trim = 0.1)
```

```
## [1] 20.22835
```

```r
median(vec)
```

```
## [1] 21.1596
```

```r
median(vec.ruid)
```

```
## [1] 25.63077
```

Comparación entre la moda, la mediana y la media

- **mean (medio)** es como el centro de masa de mi distribución
- **median (mediana)** divide en dos partes iguales
- **mode (moda)** es el valor más frecuente

Percentiles

```r
quantile(iris$Sepal.Length, seq(0,1,0.01))
```

```
##    0%    1%    2%    3%    4%    5%    6%    7%    8%    9%   10%   11%   12%
## 4.300 4.400 4.400 4.547 4.600 4.600 4.694 4.743 4.800 4.800 4.800 4.900 4.900
##   13%   14%   15%   16%   17%   18%   19%   20%   21%   22%   23%   24%   25%
## 4.900 4.900 5.000 5.000 5.000 5.000 5.000 5.000 5.029 5.100 5.100 5.100 5.100
##   26%   27%   28%   29%   30%   31%   32%   33%   34%   35%   36%   37%   38%
## 5.100 5.123 5.200 5.200 5.270 5.400 5.400 5.400 5.400 5.500 5.500 5.500 5.500
##   39%   40%   41%   42%   43%   44%   45%   46%   47%   48%   49%   50%   51%
## 5.511 5.600 5.600 5.600 5.607 5.700 5.700 5.700 5.700 5.700 5.800 5.800 5.800
##   52%   53%   54%   55%   56%   57%   58%   59%   60%   61%   62%   63%   64%
## 5.800 5.800 5.900 5.900 6.000 6.000 6.000 6.000 6.100 6.100 6.100 6.100 6.200
##   65%   66%   67%   68%   69%   70%   71%   72%   73%   74%   75%   76%   77%
## 6.200 6.234 6.300 6.300 6.300 6.300 6.300 6.328 6.400 6.400 6.400 6.400 6.473
##   78%   79%   80%   81%   82%   83%   84%   85%   86%   87%   88%   89%   90%
## 6.500 6.500 6.520 6.600 6.700 6.700 6.700 6.700 6.700 6.763 6.800 6.861 6.900
##   91%   92%   93%   94%   95%   96%   97%   98%   99%  100%
## 6.900 7.008 7.157 7.200 7.255 7.408 7.653 7.700 7.700 7.900
```

```r
quantile(iris$Sepal.Length, seq(0,1,0.25))
```

```
##   0%  25%  50%  75% 100%
##  4.3  5.1  5.8  6.4  7.9
```

Resumen del Data Frame

```r
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Análisis por especie

- Usando el comando tapply analice la media, la mediana y los cuartiles para las tres especies de Iris para las cuatro variables.

```
tapply(iris$Petal.Length, iris$Species, summary)
```

```
## $setosa
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.400   1.500   1.462   1.575   1.900
##
## $versicolor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    4.00    4.35    4.26    4.60    5.10
##
## $virginica
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.500   5.100   5.550   5.552   5.875   6.900
```

```
tapply(iris$Petal.Width, iris$Species, summary)
```

```
## $setosa
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.200   0.200   0.246   0.300   0.600
##
## $versicolor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.200   1.300   1.326   1.500   1.800
##
## $virginica
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   1.800   2.000   2.026   2.300   2.500
```

```
tapply(iris$Sepal.Length, iris$Species, summary)
```

```
## $setosa
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.300   4.800   5.000   5.006   5.200   5.800
##
## $versicolor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   5.600   5.900   5.936   6.300   7.000
##
## $virginica
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   6.225   6.500   6.588   6.900   7.900
```

```
tapply(iris$Sepal.Width, iris$Species, summary)
```

```
## $setosa
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.300   3.200   3.400   3.428   3.675   4.400
##
## $versicolor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.525   2.800   2.770   3.000   3.400
##
## $virginica
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.200   2.800   3.000   2.974   3.175   3.800
```

Medidas de Dispersión

```
max(iris$Sepal.Length) - min(iris$Sepal.Length)
```

```
## [1] 3.6
```

```
var(iris$Sepal.Length)
```

```
## [1] 0.6856935
```

```
sd(iris$Sepal.Length)
```

```
## [1] 0.8280661
```

```
aad <- function(x, fun = median){
  mean(abs(x - fun(x)))
}
```

```
aad(iris$Sepal.Length)
```

```
## [1] 0.6846667
```

```
aad(iris$Sepal.Length, mean)
```

```
## [1] 0.6875556
```

```
mad(iris$Sepal.Length)
```

```
## [1] 1.03782
```

```
IQR(iris$Sepal.Length)
```

```
## [1] 1.3
```

Estadísticas Multivariadas

```
cov(iris$Sepal.Length, iris$Sepal.Width)
```

```
## [1] -0.042434
```

```
cov(iris[,1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
## Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
## Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
## Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

```r
cor(iris[,1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

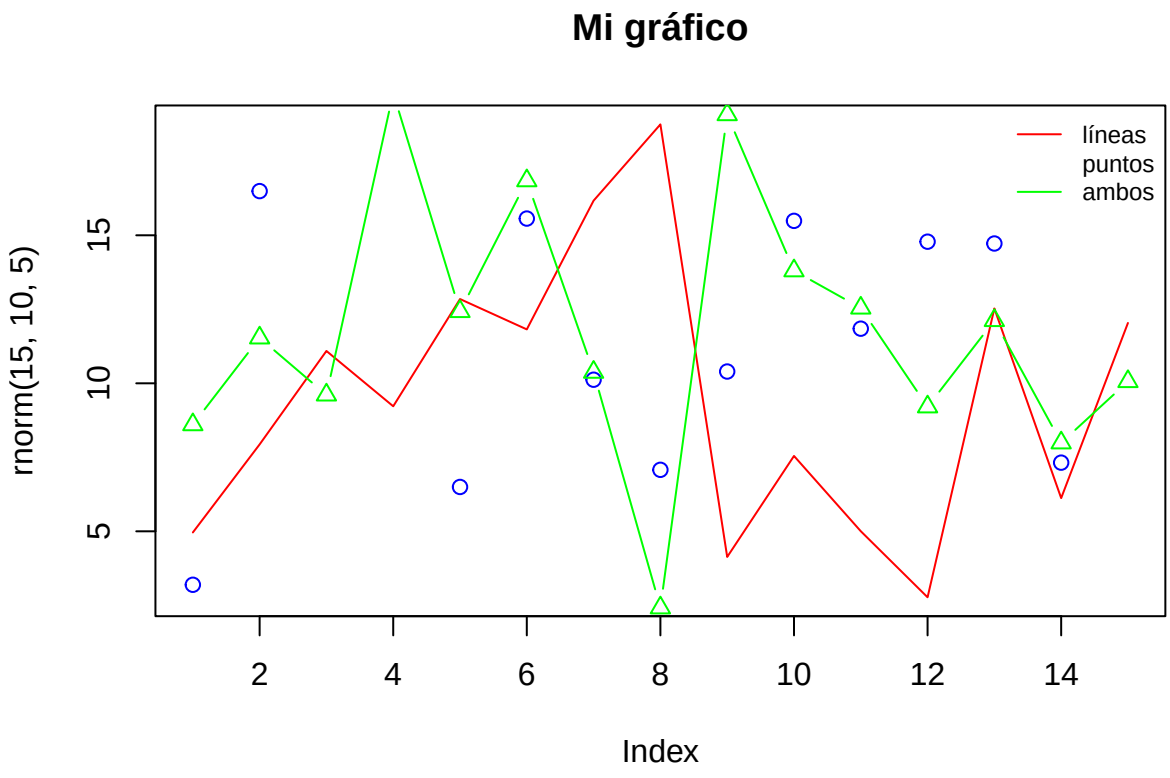
Tablas de Contingencia

```r
table(iris$Species, iris$Sepal.Length > 5)
```

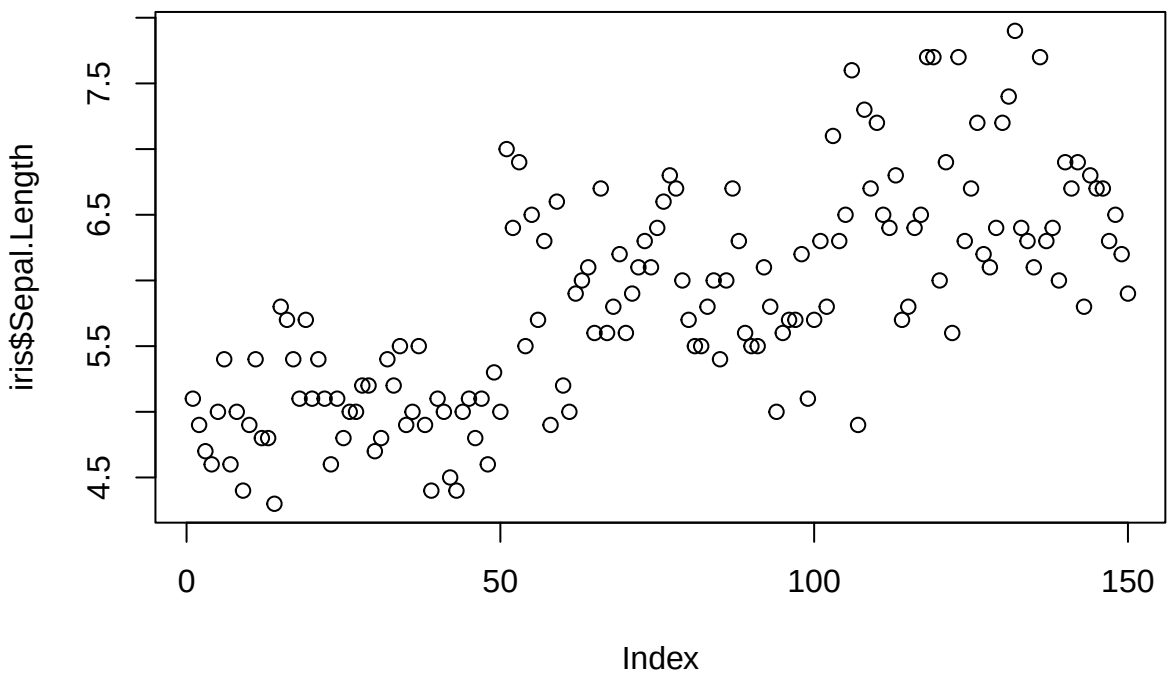
```
##
##              FALSE TRUE
##   setosa        28   22
##   versicolor     3   47
##   virginica      1   49
```

Visualización

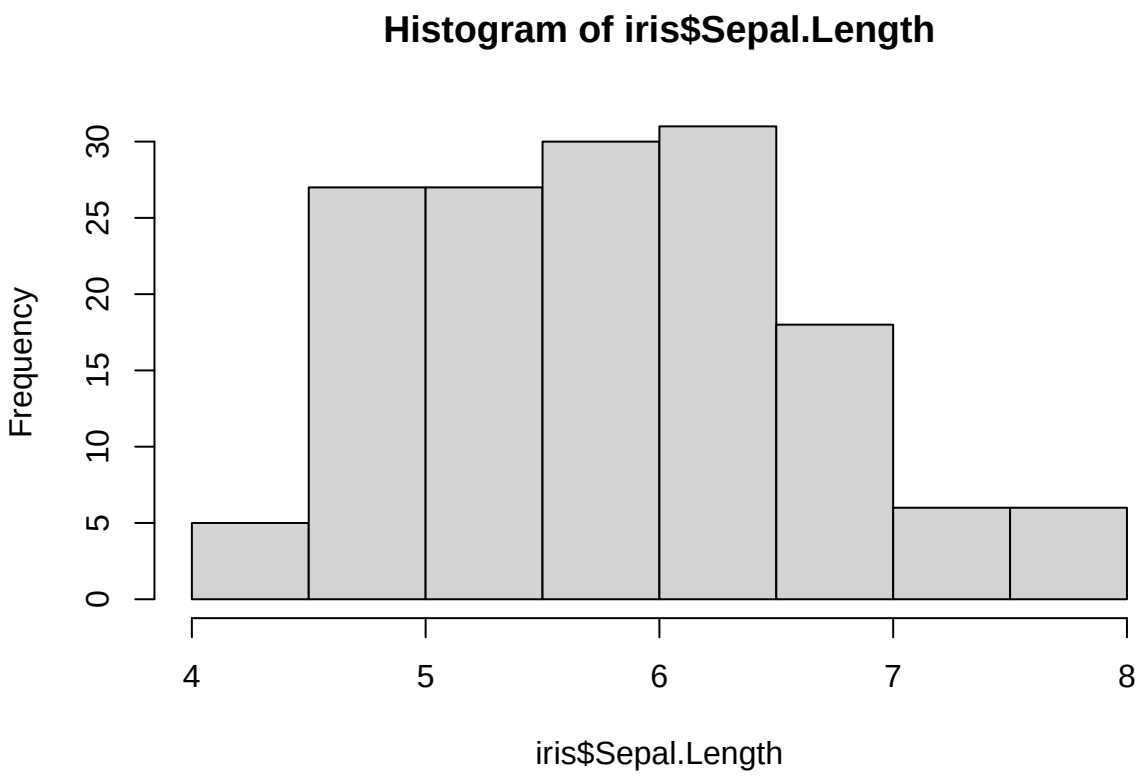
```r
plot(rnorm(15, 10, 5), col="red", type="l")
lines(rnorm(15, 10, 5), col="blue", type="p", pch=1)
lines(rnorm(15, 10, 5), col="green", type="b", pch=2)
title(main="Mi gráfico")
legend('topright', c("líneas", "puntos", "ambos"), lty=c(1, 0, 1), col=c("red", "blue", "green"), bty='
```



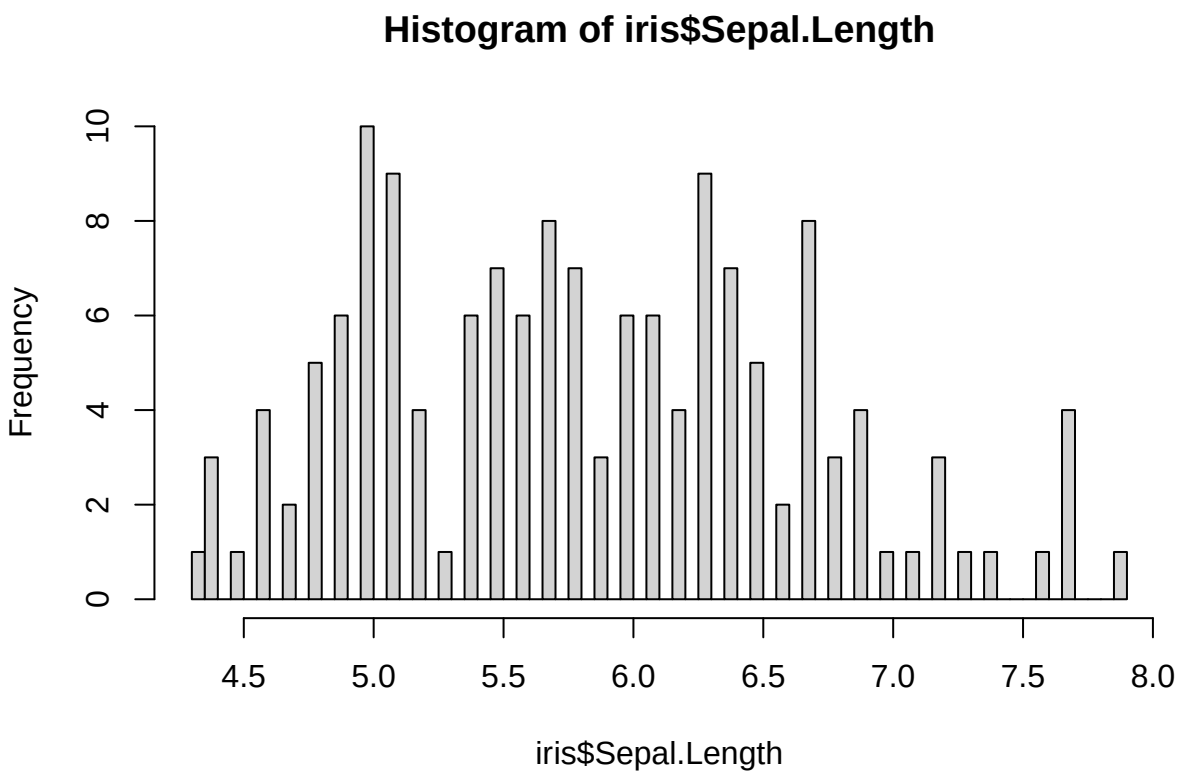
```r
plot(iris$Sepal.Length)
```




```r
hist(iris$Sepal.Length)
```



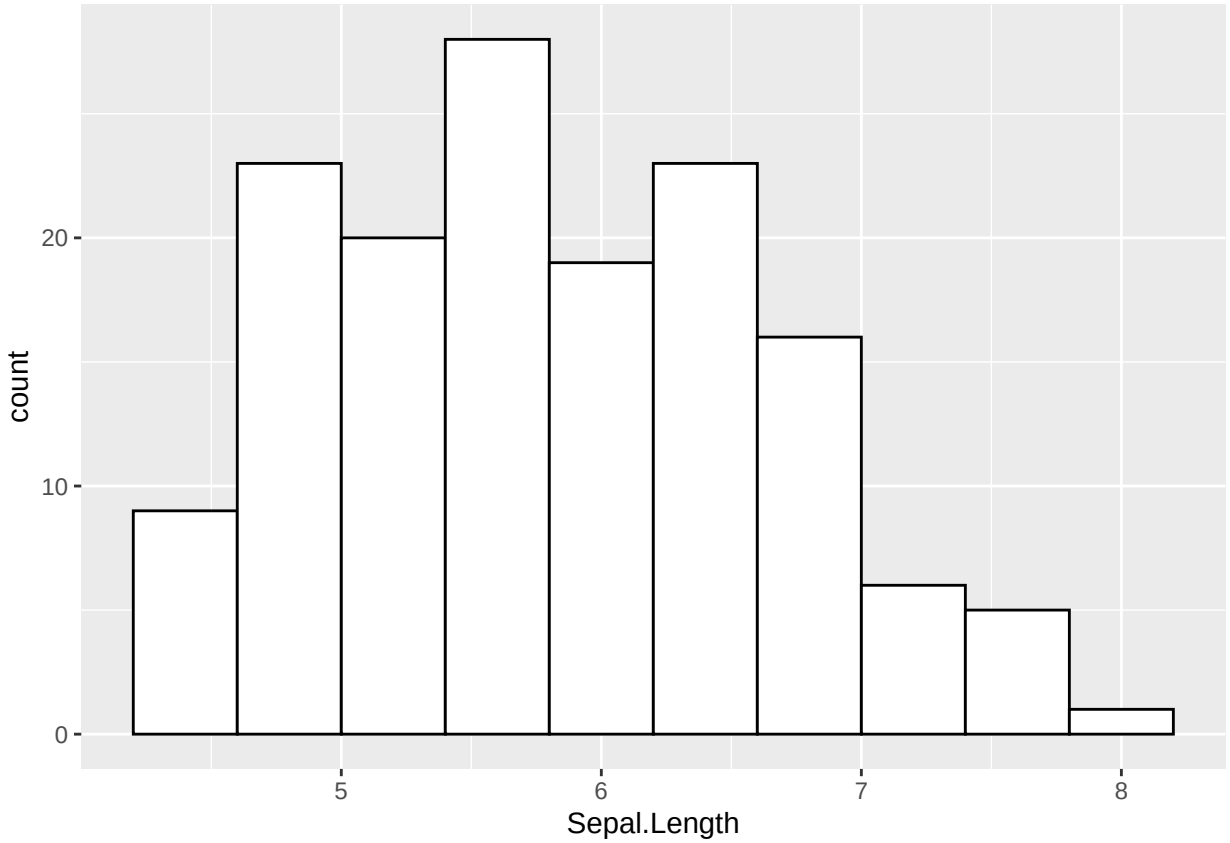

```r
hist(iris$Sepal.Length, nclass = 100)
```



ggplot2

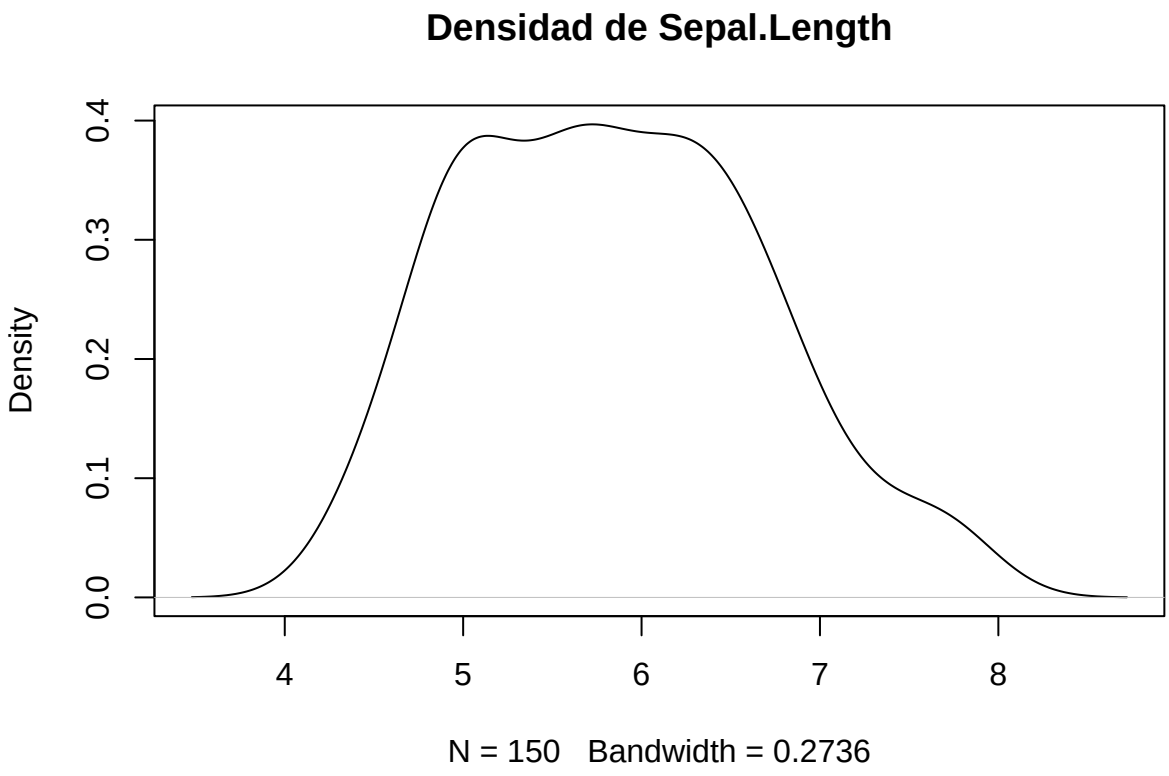
```
library(ggplot2)
```

```
ggplot(iris, aes(x = Sepal.Length)) + geom_histogram(bins = 10, color = 'black', fill = 'white')
```



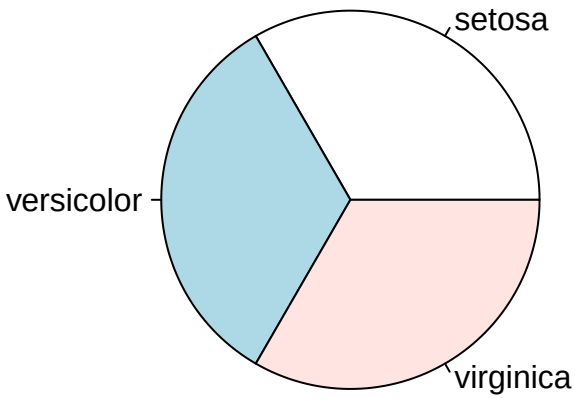
Densidad

```
plot(density(iris$Sepal.Length), main='Densidad de Sepal.Length')
```



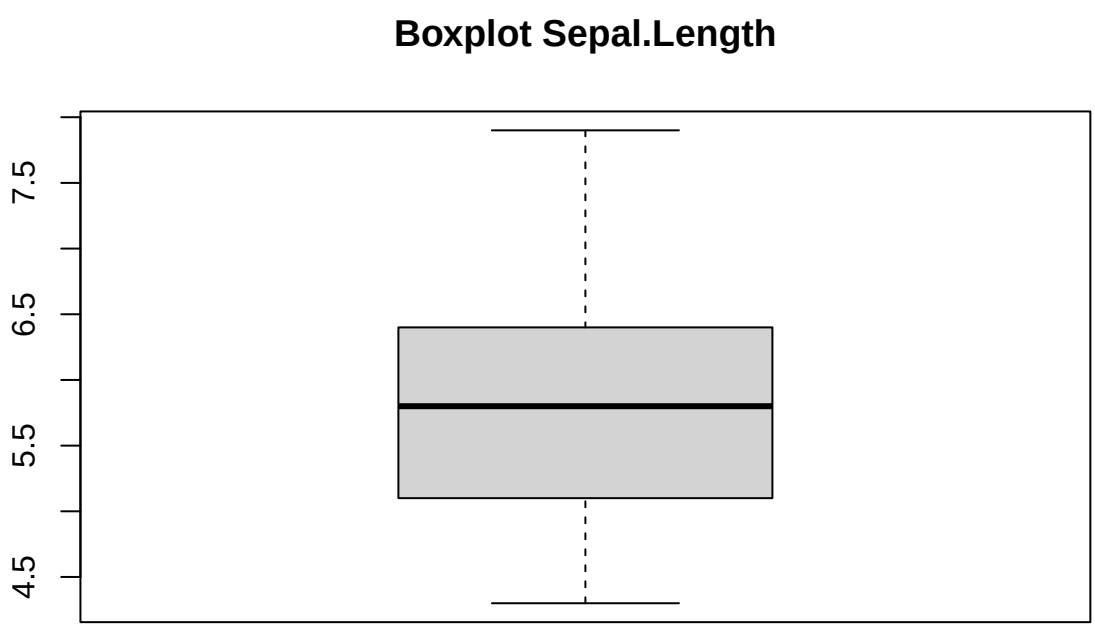
Pie Chart

```
pie(table(iris$Species))
```

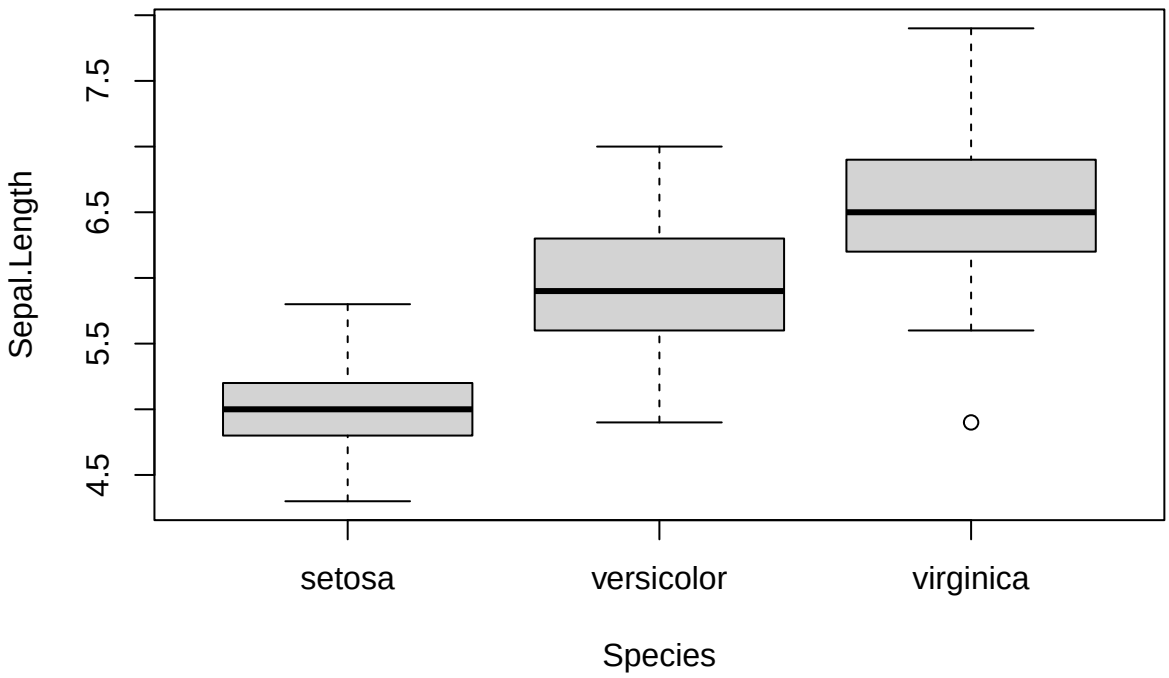


Boxplots

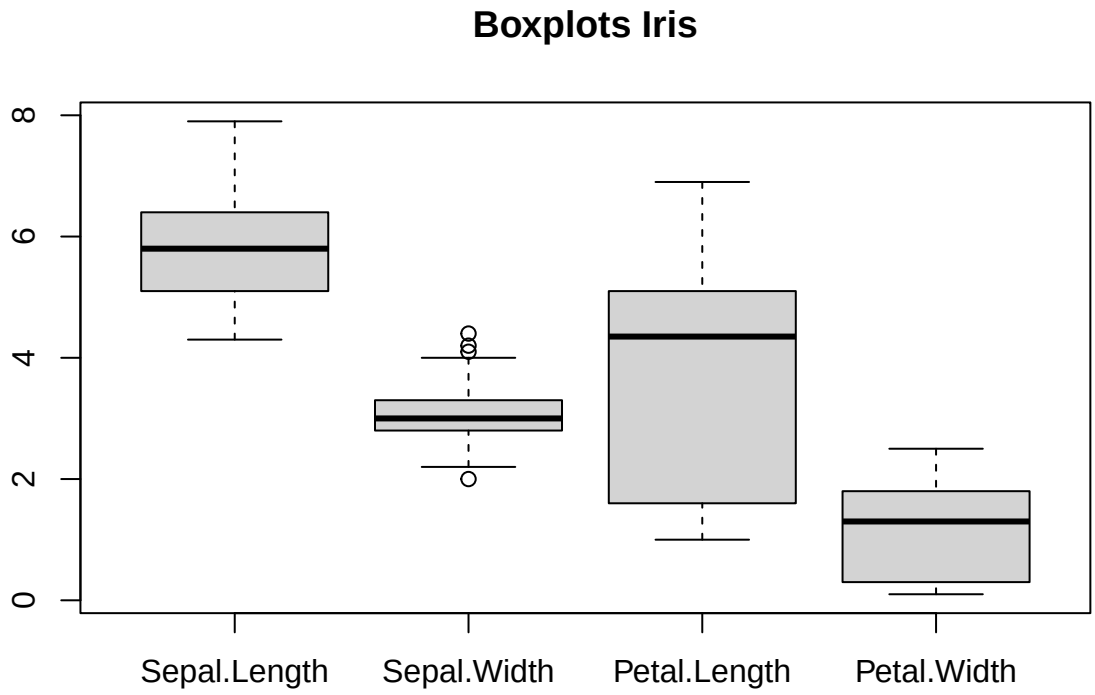
```
boxplot(iris$Sepal.Length, main='Boxplot Sepal.Length')
```



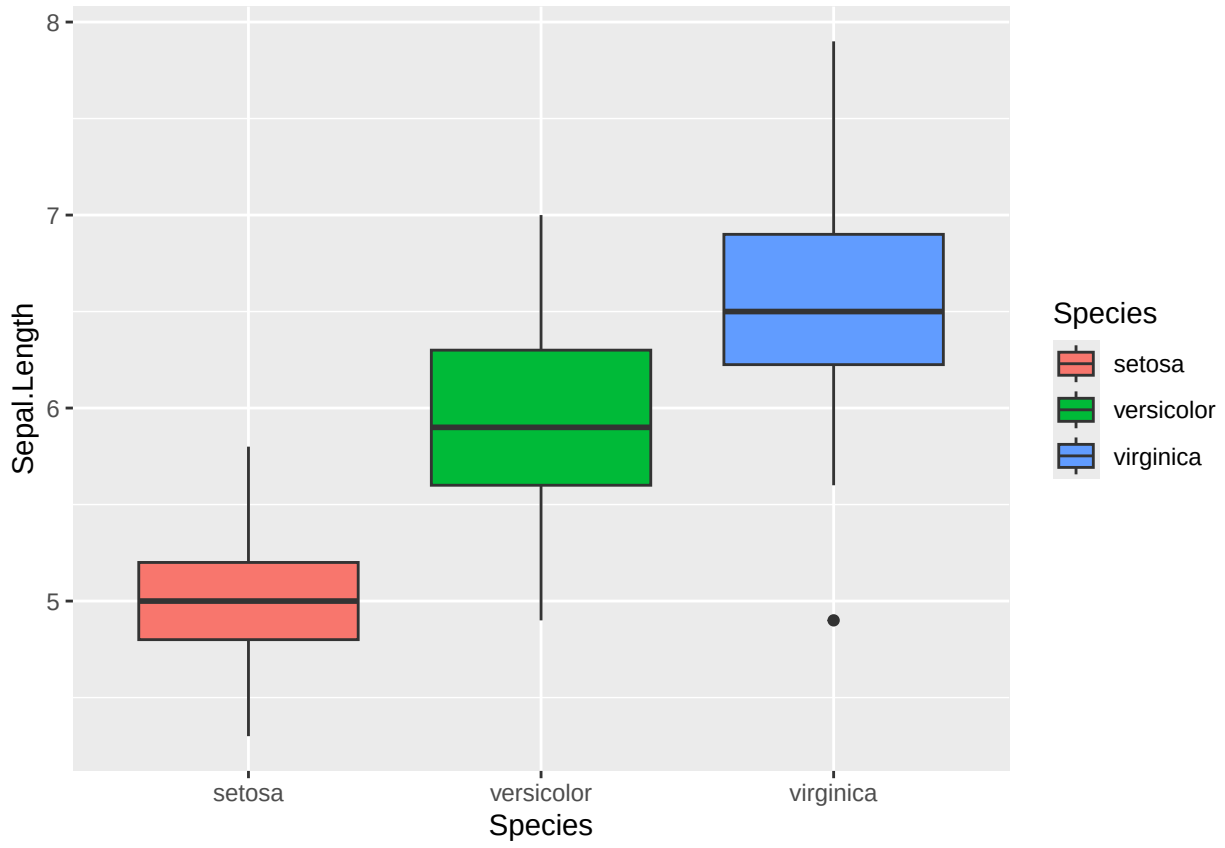

```
boxplot(Sepal.Length ~ Species, data=iris, ylab='Sepal.Length')
```




```
boxplot(iris[,1:4], main='Boxplots Iris')
```

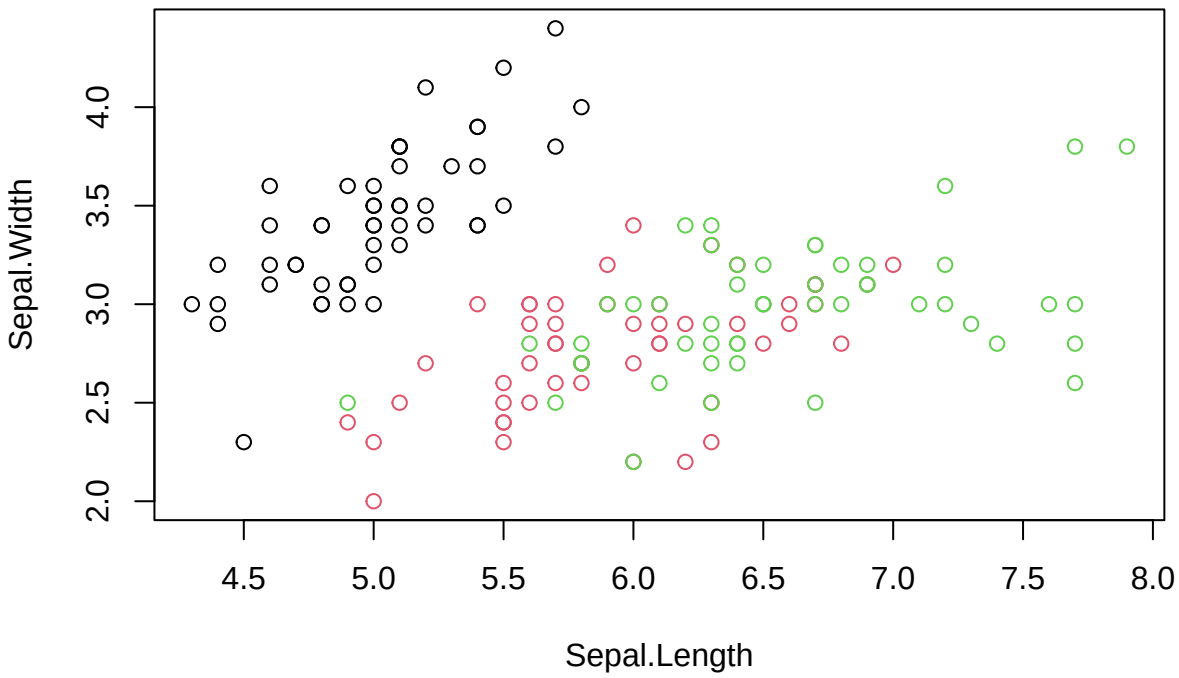



```
ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) + geom_boxplot()
```

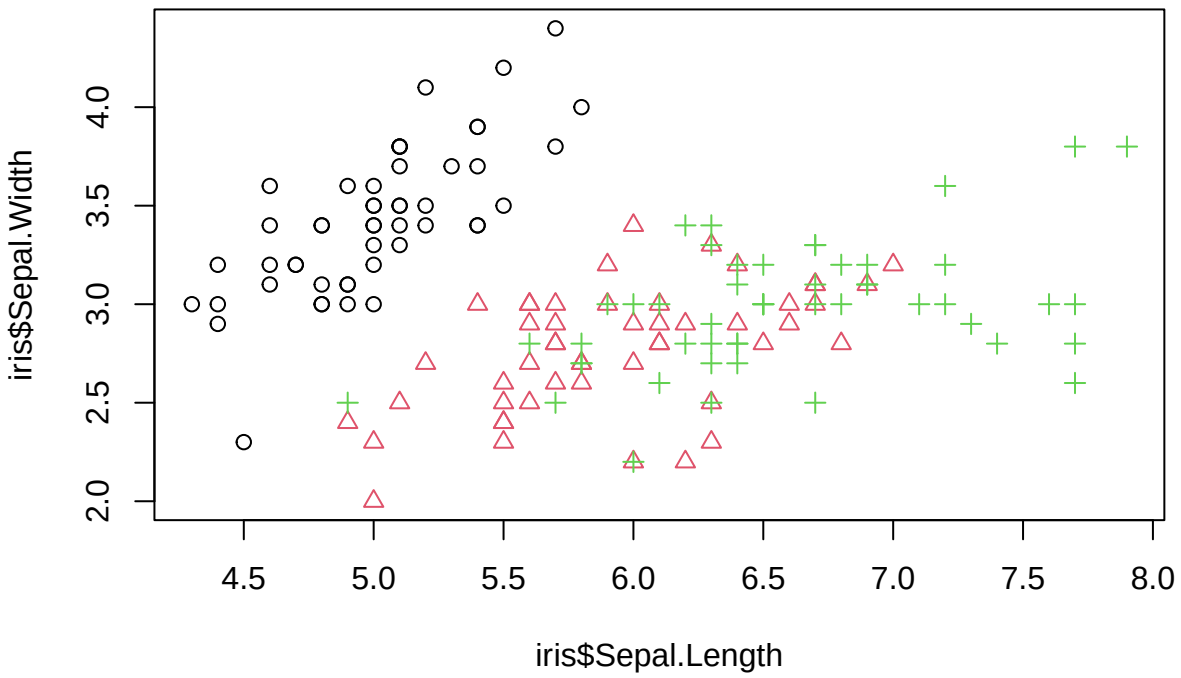


Diagramas de Disperción

```
# El ancho del sépalo vs el largo del sépalo
plot(Sepal.Width ~ Sepal.Length, col=iris$Species)
```



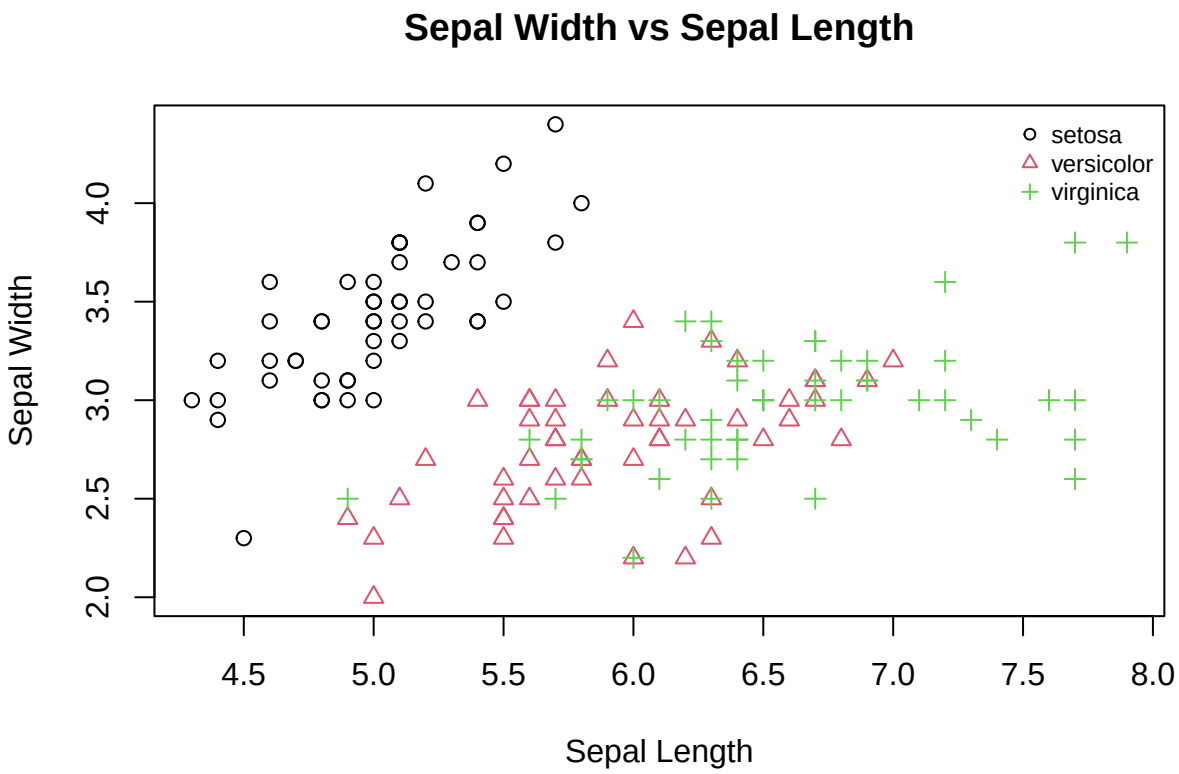

```r
# Equivalente
plot(iris$Sepal.Length, iris$Sepal.Width, col=iris$Species, pch=as.numeric(iris$Species))
```



Despues de realizar correcciones

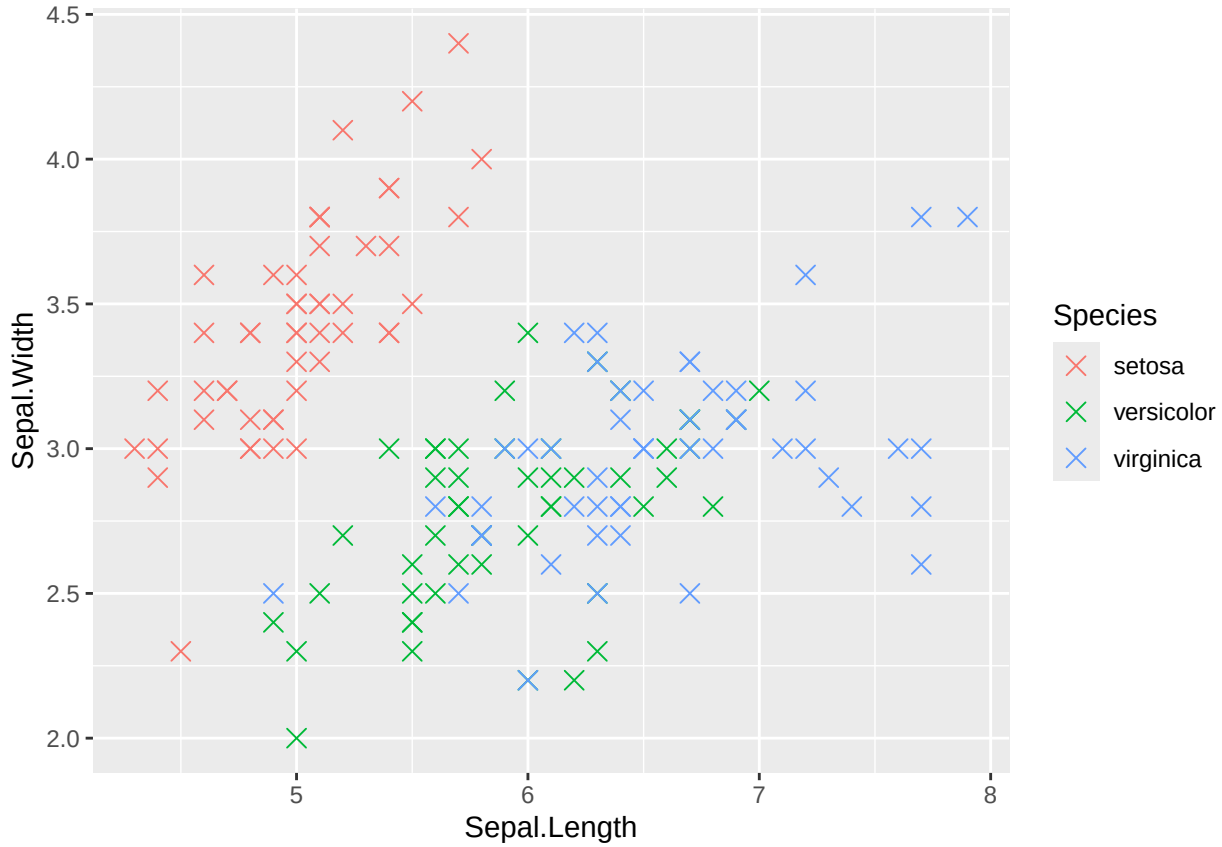
```r
# Creamos el gráfico
plot(iris$Sepal.Length, iris$Sepal.Width,
     col = as.numeric(iris$Species),
     pch = as.numeric(iris$Species),
     xlab = "Sepal Length",
     ylab = "Sepal Width",
     main = "Sepal Width vs Sepal Length")

# Agregamos la leyenda
legend("topright",
       legend = levels(iris$Species),
       col = 1:3,
       pch = 1:3,
       bty = "n",
       cex = 0.75)
```



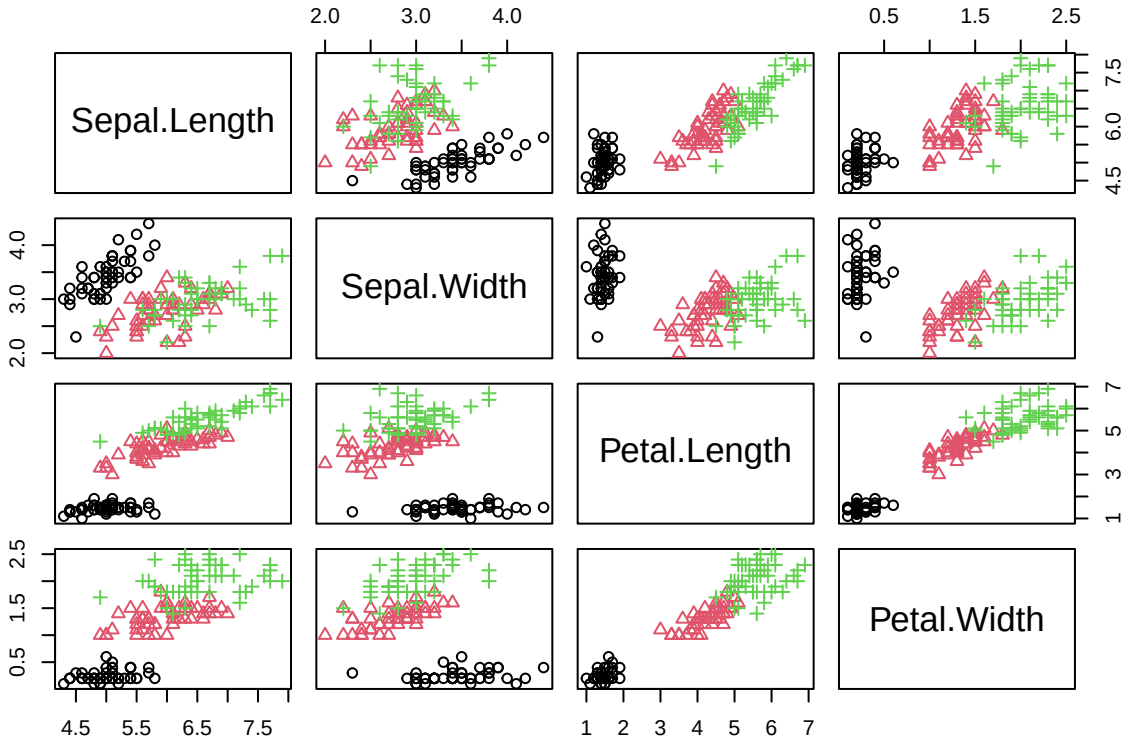
Mismo grafico pero con ggplot

```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) + geom_point(size=3, shape=4)
```



Ahora grafiquemos todos los pares de las 4 variables del dataset iris usando un color y un carácter distinto para cada especie:

```
pairs(iris[,1:4], pch=as.numeric(iris$Species), col=iris$Species)
```

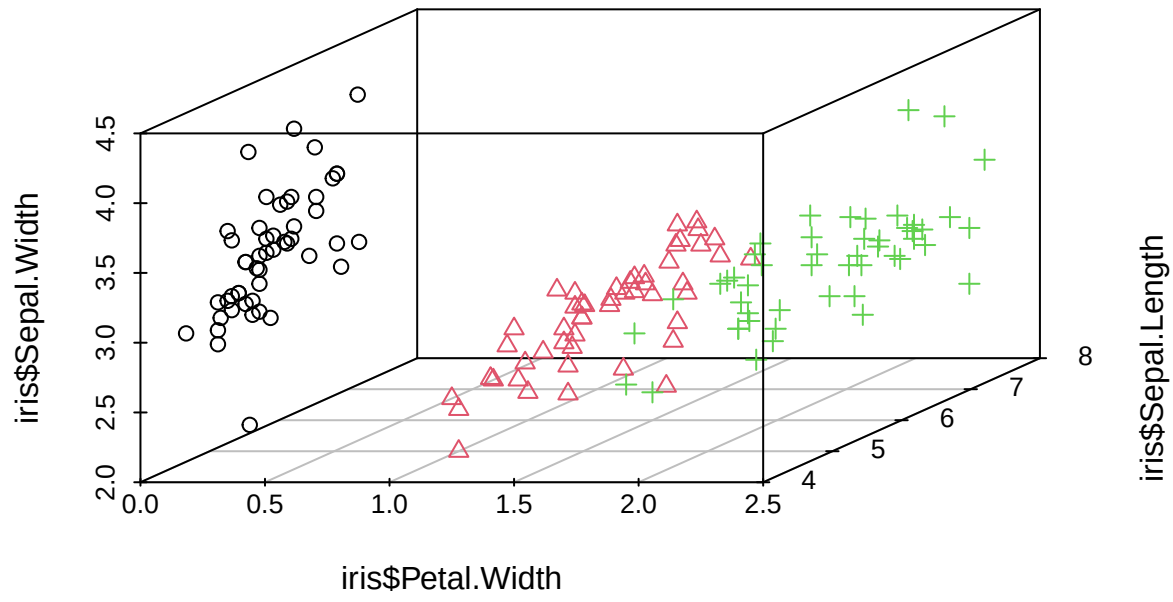


Dispercion en 3d

```
install.packages("scatterplot3d",dependencies=T)
```

```
## Installing package into '/home/gaston/R/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

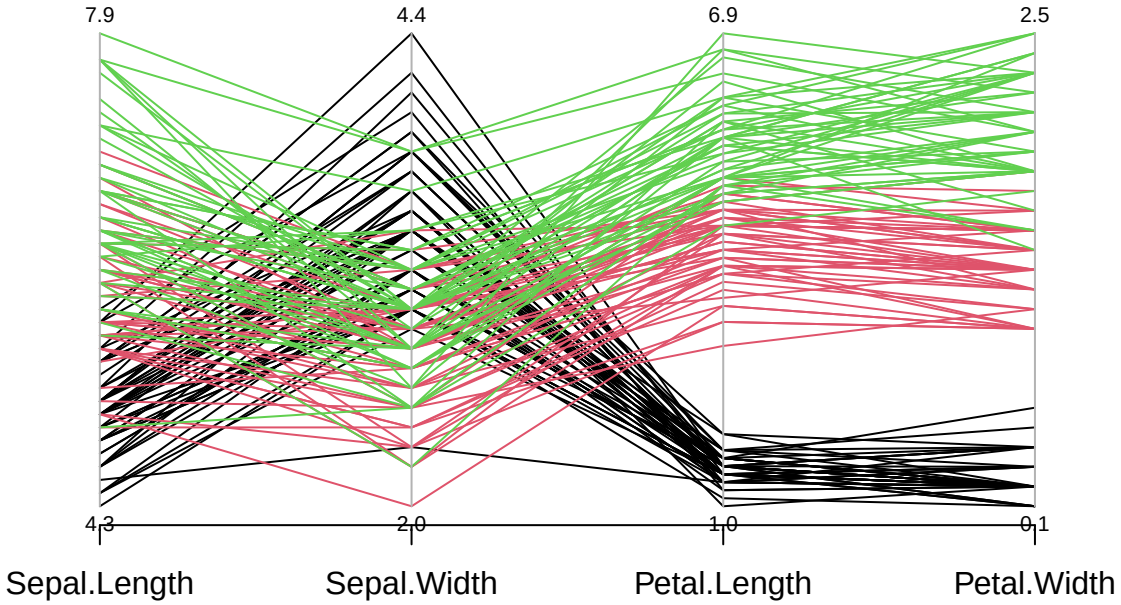
```
library(scatterplot3d)
scatterplot3d(iris$Petal.Width, iris$Sepal.Length,
iris$Sepal.Width, color=as.numeric(iris$Species),
pch=as.numeric(iris$Species))
```



Gráficos de Coordenadas Paralelas

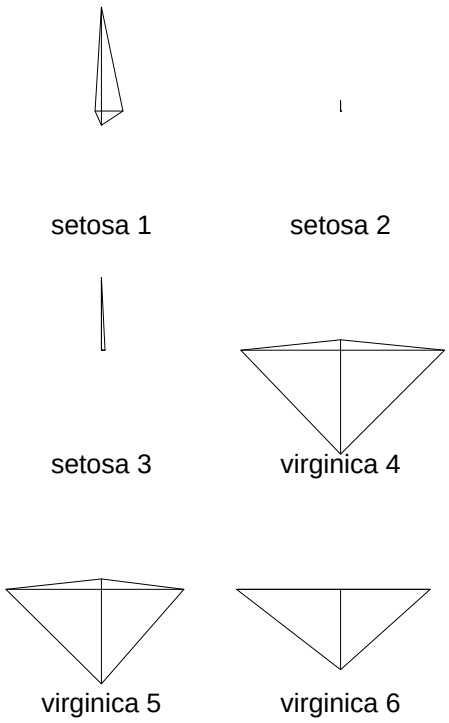
```
library(MASS)
```

```
parcoord(iris[1:4], col=iris$Species, var.label=TRUE)
```



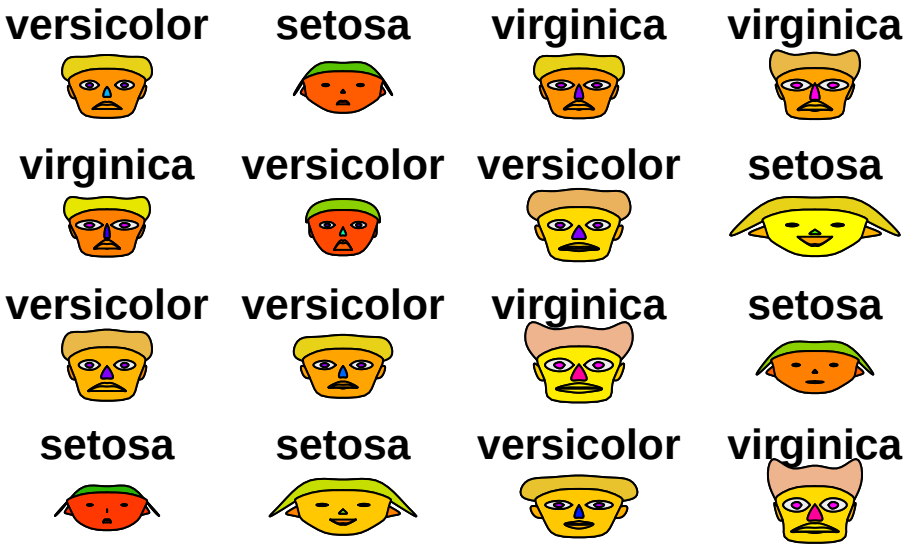
Gráficos de Estrellas

```
iris_sample1<-iris[sample(1:dim(iris)[1],size=6,replace=F),]
rownames(iris_sample1)<-
paste(as.character(iris_sample1$Species),1:6)
stars(iris_sample1[1:4])
```



Caras de Chernoff

```
library("aplpack")
iris_sample<-iris[sample(1:dim(iris)[1],size=16,replace=F),]
faces(iris_sample[1:4],face.type=1,labels=iris_sample$Species)
```




```
## effect of variables:
##  modified item       Var
##  "height of face   " "Sepal.Length"
##  "width of face    " "Sepal.Width"
##  "structure of face" "Petal.Length"
##  "height of mouth  " "Petal.Width"
##  "width of mouth   " "Sepal.Length"
##  "smiling          " "Sepal.Width"
##  "height of eyes   " "Petal.Length"
##  "width of eyes    " "Petal.Width"
##  "height of hair   " "Sepal.Length"
##  "width of hair    "  "Sepal.Width"
##  "style of hair    "  "Petal.Length"
##  "height of nose   "  "Petal.Width"
##  "width of nose    "  "Sepal.Length"
##  "width of ear     "  "Sepal.Width"
##  "height of ear    "  "Petal.Length"
```