***SESION 5***

***Sesión 5: Graficas con rstudio***

**Contenido**

Graficas de barras , graficas de dispersión ,etc.

**Actividad 5:**

Aplicaciones en R utilizando los datos de esta sesión.

**Histograma**

Se utilizan usando la función *hist*. Sintaxis: *hist(x,…)* Donde X es un vector.

**Histograma de Frecuencias**

Realizamos un ejemplo cargando la siguiente base de datos:

265205263307220268260234299215197286274243231267281265214318349317242258276300208187264271280242260321228250299258267293265254281294223260308235283277200235246328296276264269235290221174248263231334280265272283265262271245301280274253287258261248260274337250278254274275278250265270298257210280269251265197349280265200221265261278205286317242254235174262248250263274242260281246248271260265307243258321294328263245274270220231276228223296231301337298268267300250260276334280250257260281208299308264280274278210234265187258235269265253254280299214264267283235272287274269215318271293277290283258275251

Cargamos la base de datos de la siguiente forma:

datos<-**c**(265, 197, 349, 280, 265, 200, 221, 265, 261, 278,

205, 286, 317, 242, 254, 235, 174, 262, 248, 250,

263, 274, 242, 260, 281, 246, 248, 271, 260, 265,

307, 243, 258, 321, 294, 328, 263, 245, 274, 270,

220, 231, 276, 228, 223, 296, 231, 301, 337, 298,

268, 267, 300, 250, 260, 276, 334, 280, 250, 257,

260, 281, 208, 299, 308, 264, 280, 274, 278, 210,

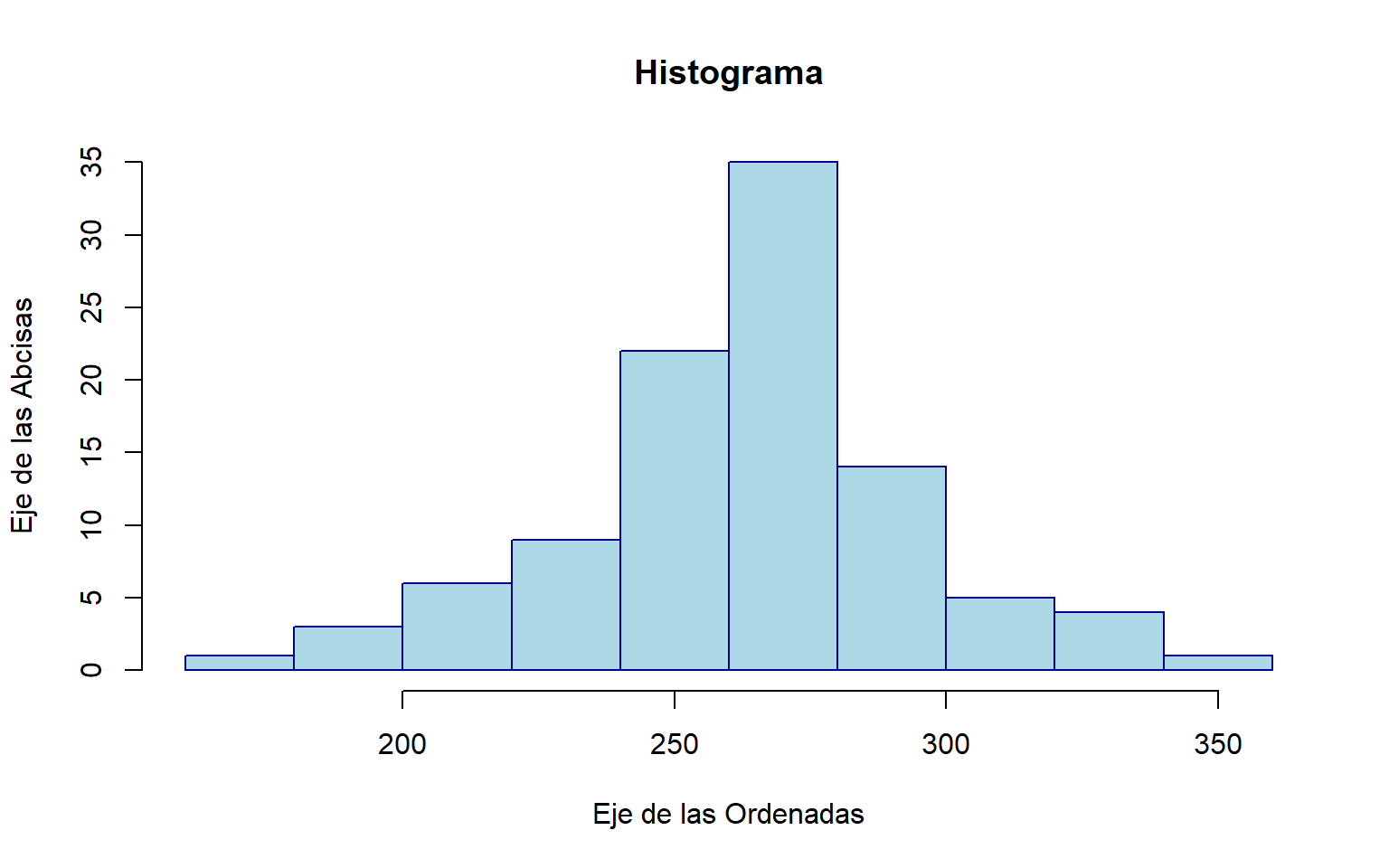
234, 265, 187, 258, 235, 269, 265, 253, 254, 280,

299, 214, 264, 267, 283, 235, 272, 287, 274, 269,

215, 318, 271, 293, 277, 290, 283, 258, 275, 251)

Y luego graficamos:

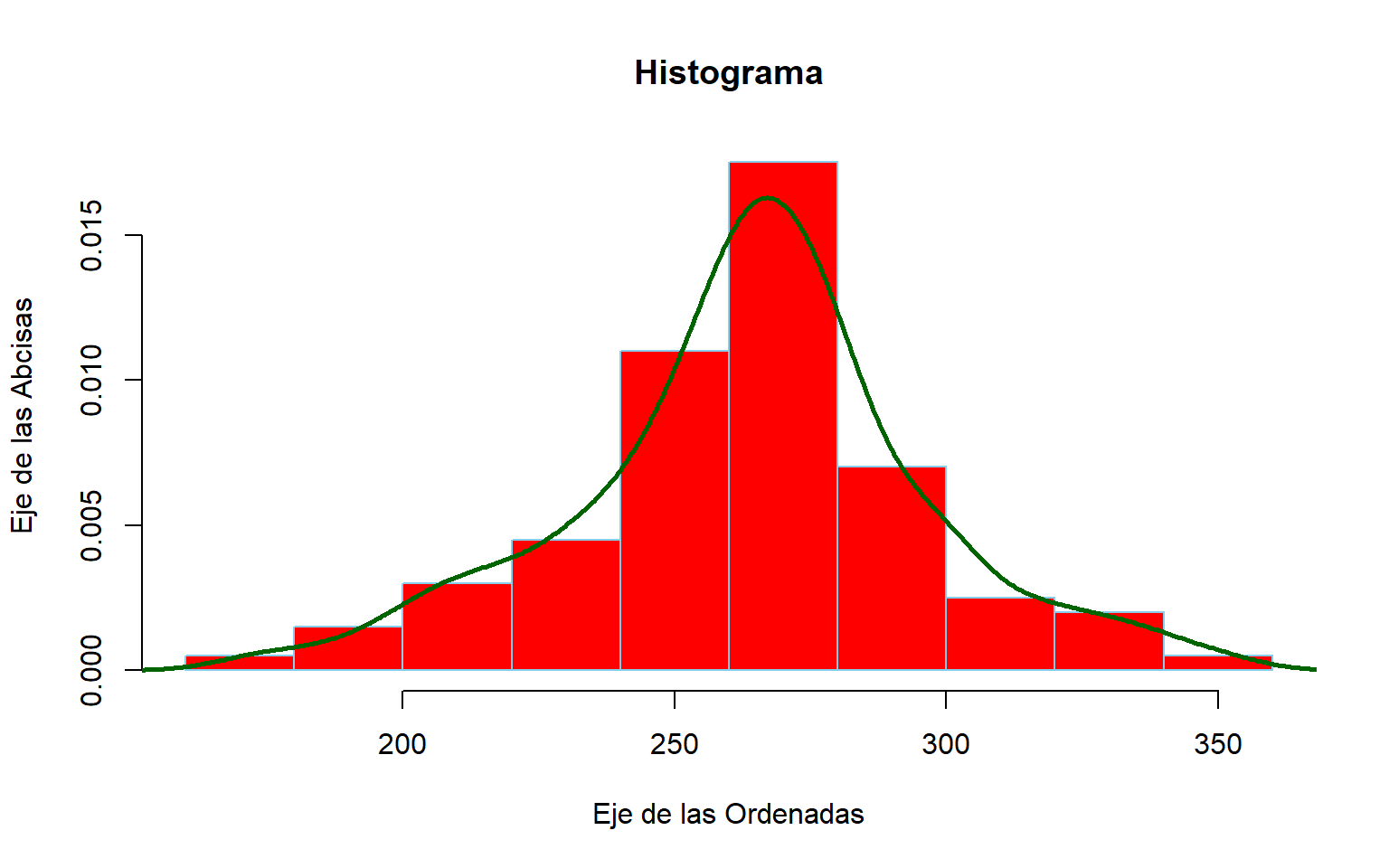
**hist**(datos, main="Histograma",col="lightblue",border="darkblue",ylab = "Eje de las Abcisas", xlab = "Eje de las Ordenadas")

 ### Histograma de Densidad

La función *hist()* por defecto, muestra el gráfico de frecuencia; también genera la representación gráfica de las densidades. Sintaxis: hist(x,prob=TRUE). Donde “TRUE” es para mostrar la densidad y no las frecuencias, y “lwd” es el grosor de la línea.

**hist**(datos, main="Histograma",col="red",border="skyblue",ylab = "Eje de las Abcisas", xlab = "Eje de las Ordenadas",prob=TRUE)

**lines**(**density**(datos),lwd=2.5,col="darkgreen")



**Diagrama de Barras**

Se realizan utilizando la función *barplot()*. Donde X puede ser un vector o matriz en la sintaxis: barplot(x,..)

Ejemplo: Los siguientes datos corresponden a las puntuaciones de un conductor de Cabify y con ello debemos realizar un gráfico de barras para frecuencias absolutas y relativas.

355543543445544555344555354555454155344355534134545544555555545555455445235535555444455355455353541335155555544255445444454555554533445545543354553354545454545445455555555545

Cargamos los datos y luego usamos la función *table* ya que toda tabla de frecuencias se denomina con dicha función.

datos<-**c**(3, 5, 5, 4, 5, 5, 3, 5, 3, 5, 4, 1, 3, 3, 5, 1, 5, 5, 5, 5, 5, 5, 4, 4, 2, 5, 5, 4, 4,

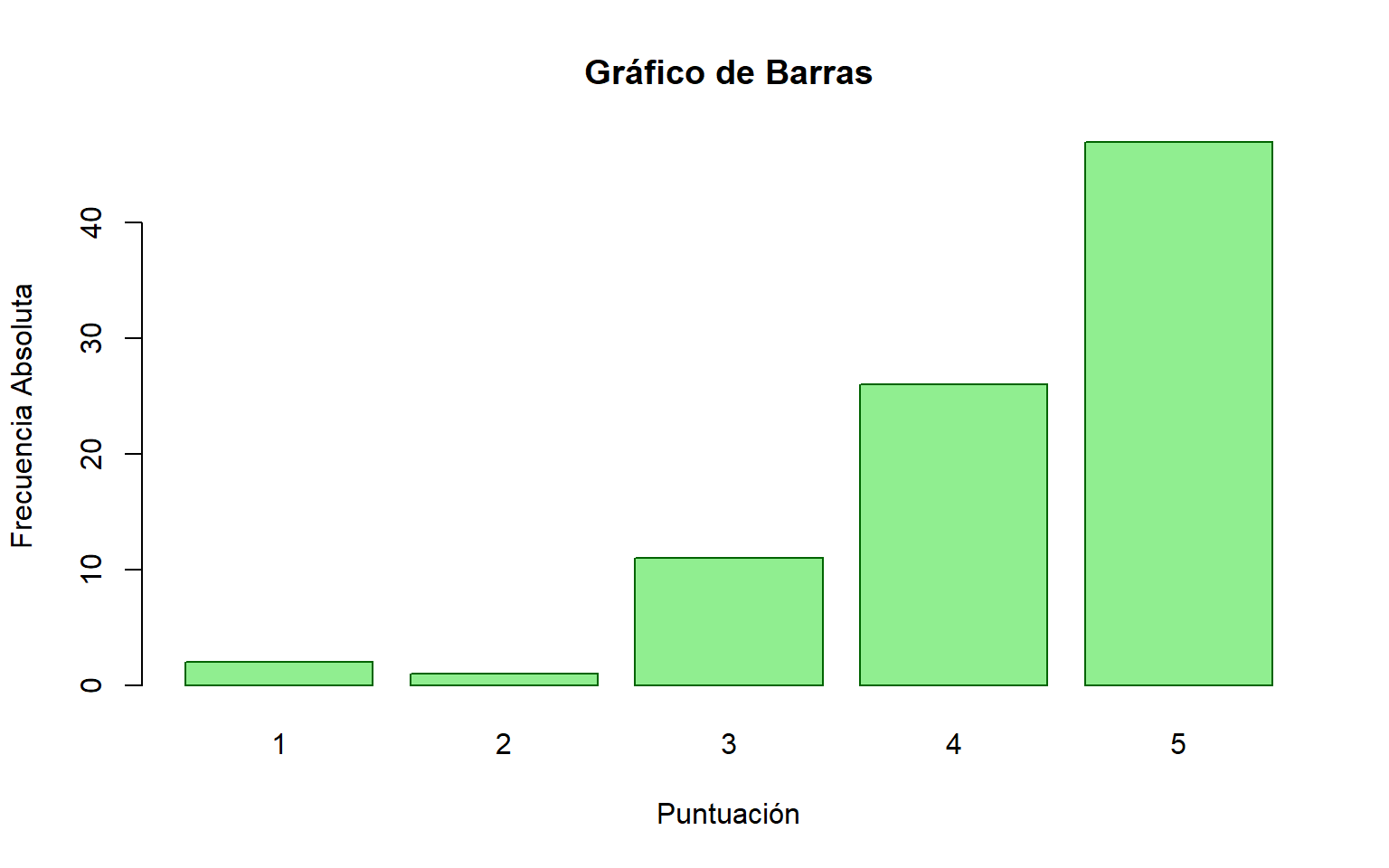
5, 4, 4, 4, 4, 5, 4, 5, 5, 5, 5, 5, 4, 5, 3, 3, 4, 4, 5, 5, 4, 5, 5, 4, 3, 3, 5, 4, 5,

5, 3, 3, 5, 4, 5, 4, 5, 4, 5, 4, 5, 4, 5, 4, 4, 5, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 5)

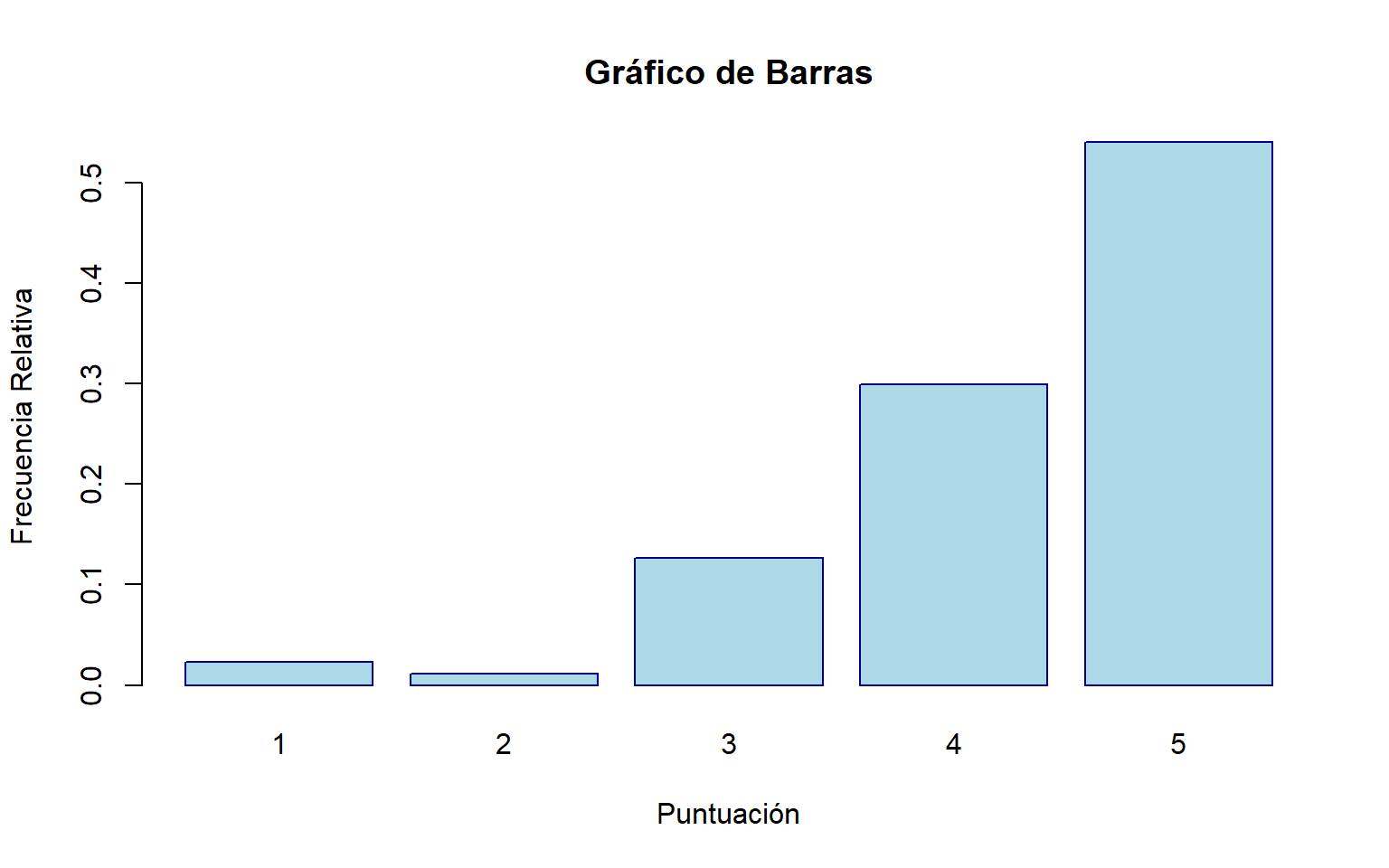
tabla<-**table**(datos)

Realizamos el gráfico de frecuencia absoluta con los siguientes comandos:

**barplot**(tabla, main = "Gráfico de Barras", xlab = "Puntuación", ylab = "Frecuencia Absoluta", col = "lightgreen", border = "darkgreen")

 Finalmente, realizamos el gráfico de frecuencias relativas con los siguientes comandos: Recordar que una frecuencia relativa es una proporción de la frecuencia absoluta entre la cantidad de datos, es decir, un cociente, por ello usamos *prop*.

**barplot**(**prop.table**(tabla), main = "Gráfico de Barras", xlab = "Puntuación", ylab = "Frecuencia Relativa", col ="lightblue", border = "darkblue")



**Gráficos Circulares**

**Gráfico 2D:**

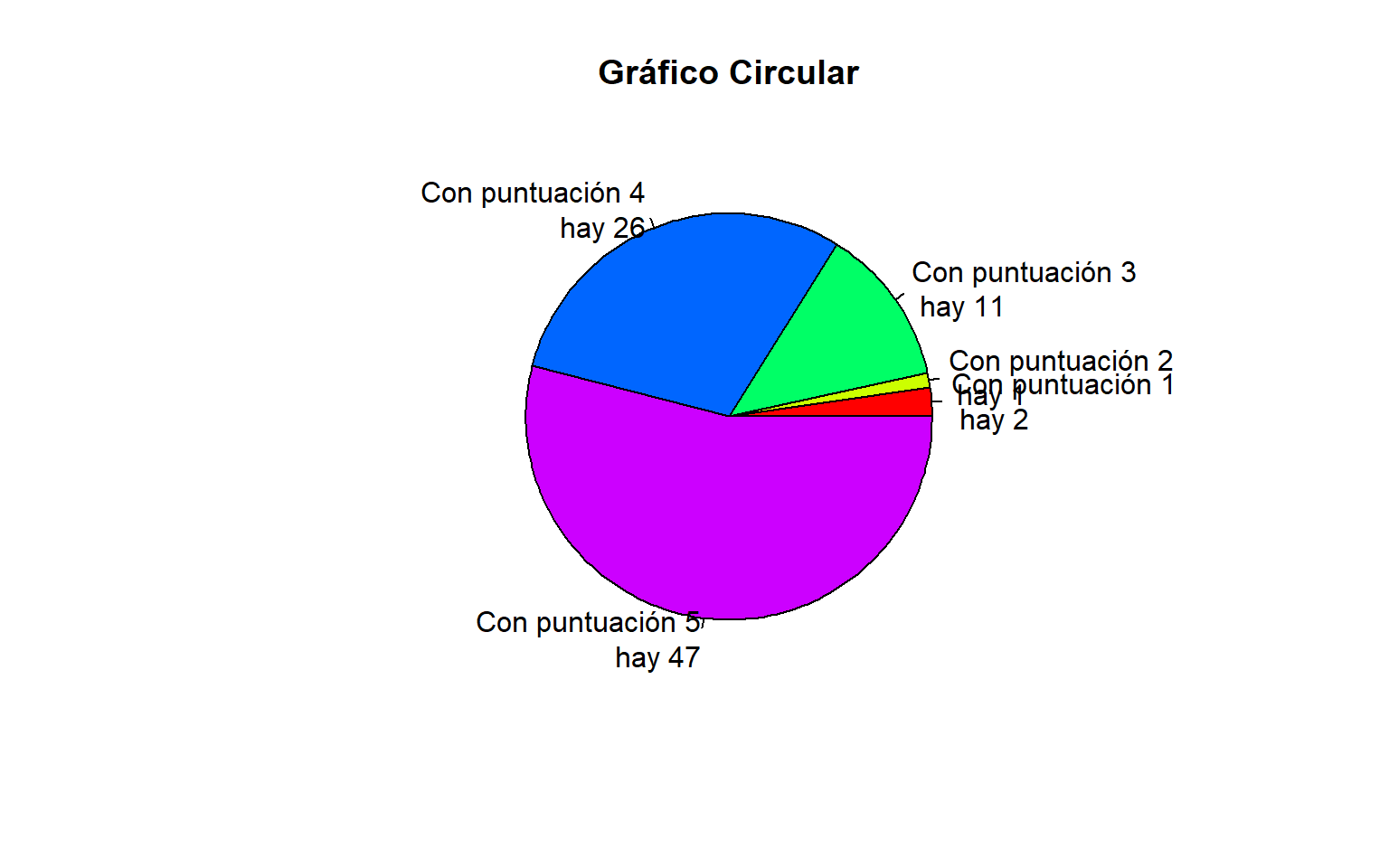
La sintaxis para realizar este tipo de gráficos es *pie(x,…)*. Donde X es un vector numérico no negativo.

Primero se debe generar las etiquetas

etiquetas<-**paste**("Con puntuación ",**names**(tabla),"\n hay ", tabla, sep="")

Ponemos que sea colorido con rainbows

**pie**(tabla, main = "Gráfico Circular", col = **rainbow**(**length**(tabla)), labels = etiquetas)



**Gráfico 3D:**

Para ello debemos installar el paquete *plotrix* con:

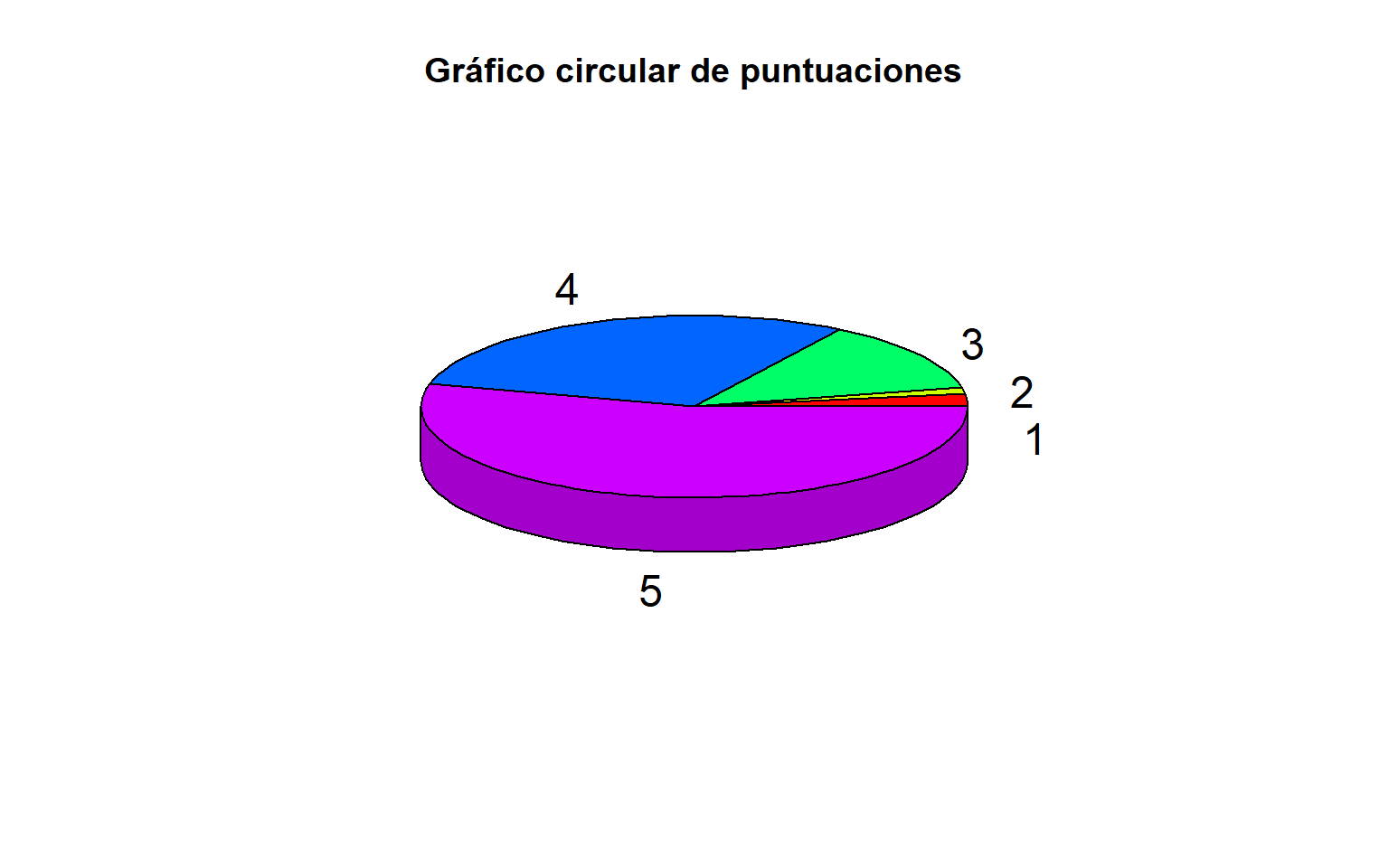
*install.packages(plotrix)*

*library(plotrix)*

Ahora usamos la función para hacer el gráfico

**library**(plotrix)

**pie3D**(tabla,main="Gráfico circular de puntuaciones",labels = **names**(tabla),col = **rainbow**(**length**(tabla)))



**Diagrama de Pareto**

La sintaxis es la siguiente: \_pareto.chart(x,…). Donde X es un vector, pero antes debemos instalar el siguiente paquete:

*install.packages(qcc)*

*library(qcc)*

Para el ejemplo se seguirá utilizando la base de datos del conductor de Cabify. Para ello, crearemos una tabla utilizando la tabla creada anteriormente y 2l *ylab2* será el porcentaje acumulado, siempre un 20% y 80%.

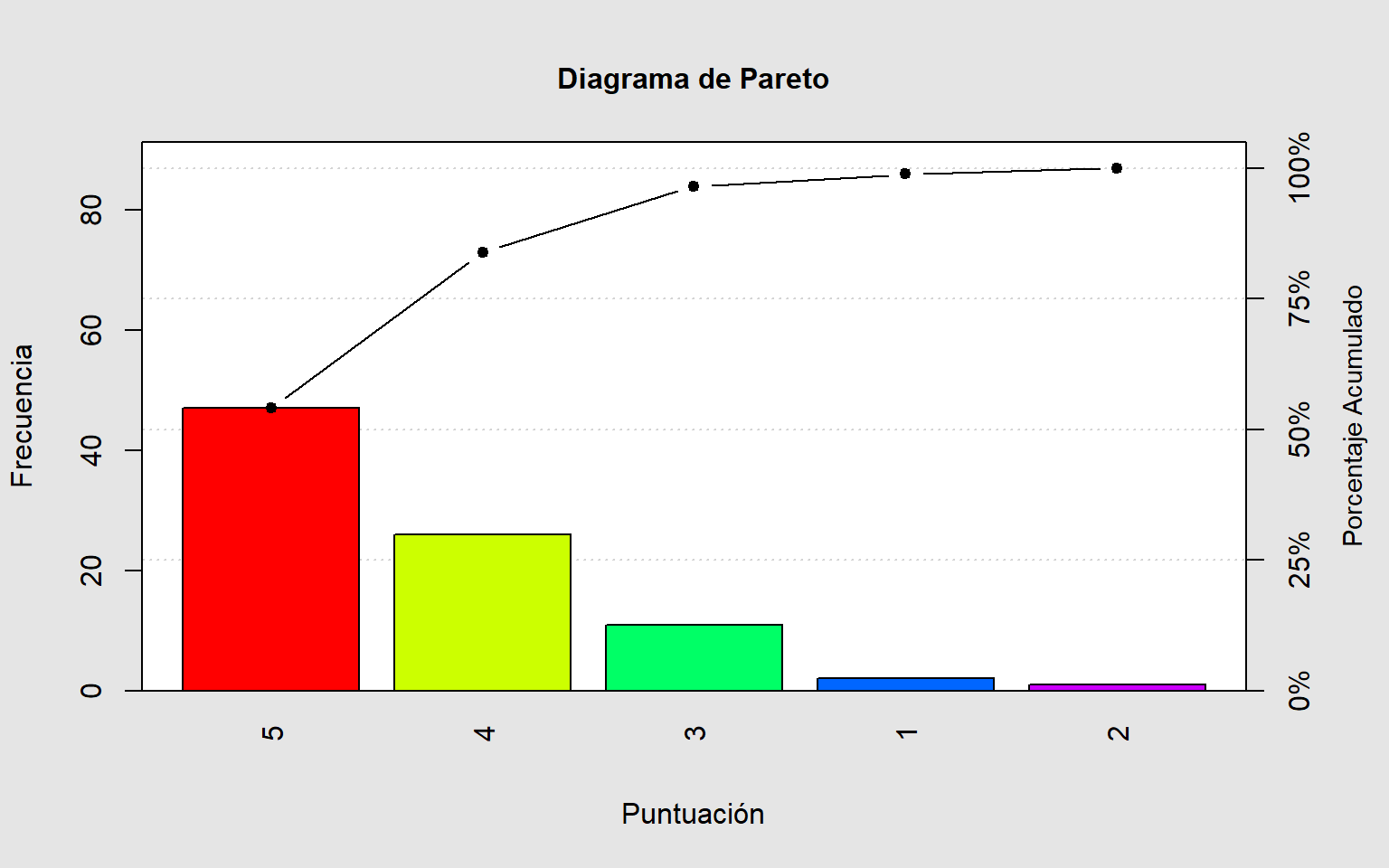
**library**(qcc)

## Warning: package 'qcc' was built under R version 4.0.2

## Package 'qcc' version 2.7

## Type 'citation("qcc")' for citing this R package in publications.

Tabla<-**pareto.chart**(tabla, main="Diagrama de Pareto", xlab="Puntuación", ylab="Frecuencia", ylab2="Porcentaje Acumulado",col=rainbow(length(tabla)))

 Para ver los datos de la Tabla creada.

Tabla

##

## Pareto chart analysis for tabla

## Frequency Cum.Freq. Percentage Cum.Percent.

## 5 47.000000 47.000000 54.022989 54.022989

## 4 26.000000 73.000000 29.885057 83.908046

## 3 11.000000 84.000000 12.643678 96.551724

## 1 2.000000 86.000000 2.298851 98.850575

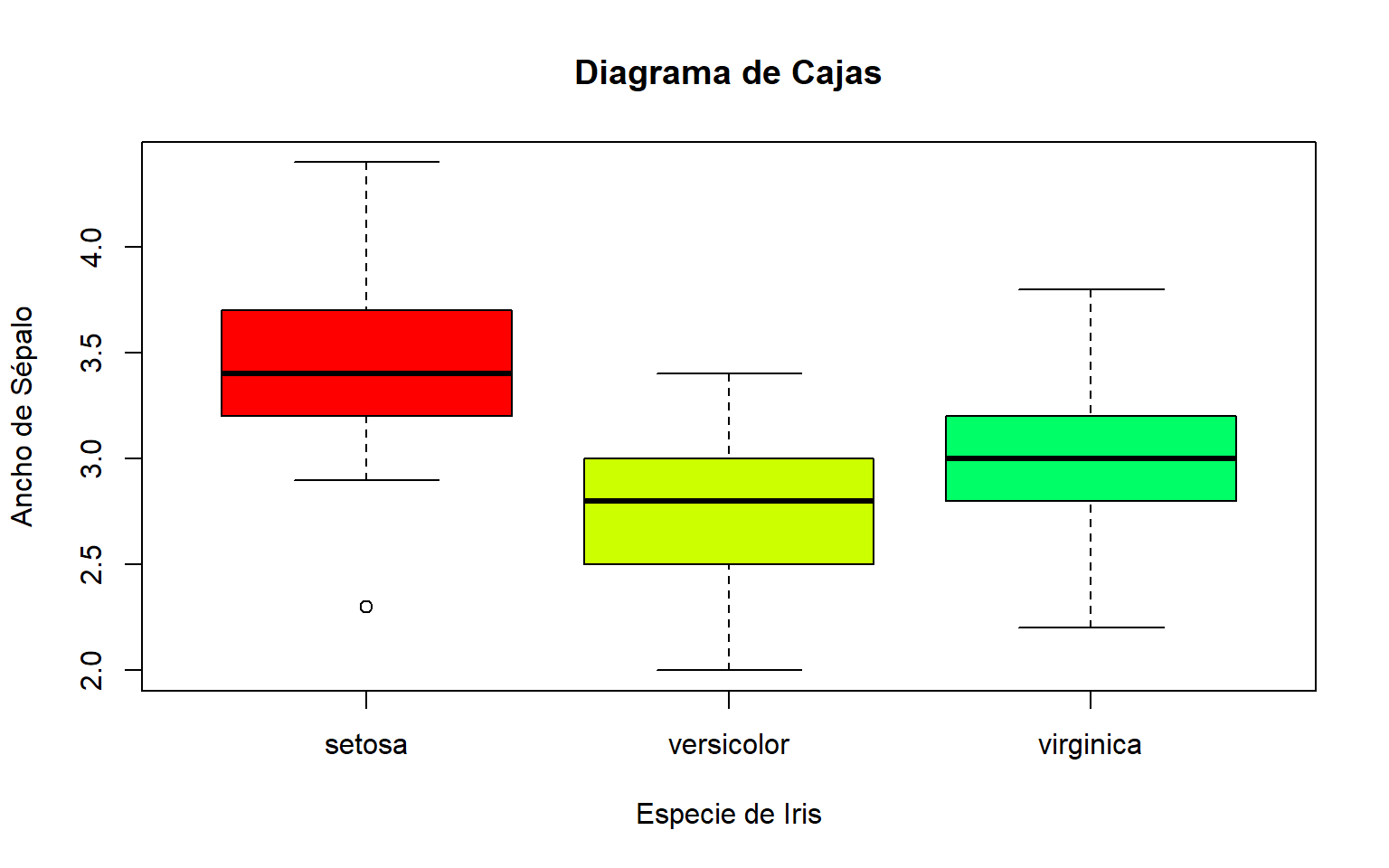
## 2 1.000000 87.000000 1.149425 100.000000

**Diagrama de Cajas**

La sintaxis es la siguiente: *boxplot(x~grupo)*. Donde X es un vector numérico que debe dividirse en grupos en referencia a la variable grupoy este último debe ser cualitativo.

Para el ejmeplo se usará la data *iris* que viene instalada en R, utilizando los siguientes comandos para obtener un diagrama de cajasusando el ancho de sépalos de la data *iris* con agrupamiento de especies:

**boxplot**(iris$Sepal.Width ~ iris$Species, col=**rainbow**(**length**(tabla)), main="Diagrama de Cajas", xlab = "Especie de Iris", ylab = "Ancho de Sépalo")

 Del gráfico se desprende que, la mediana de las variables (iris setosa, versicolor y virgínica) no son homogéneas. Dado que la iris setosa presenta una mediana mayor al de los otros tipos de iris. Además, en la iris setosa se puede observar un dato atípico. Se recomienda analizarlo.

**Diagrama de Tallos y Hojas**

La sintaxis es la siguiente: *stem(x,…)*. Donde X es un vector numérico.

Ejemplo: Realizar un diagrama de tallos y hojas para 20 plantas, extraídos de manera aleatoria de la data iris, en referencia al largo del sépalo. Primero extraemos los datos:

dato<-iris[1:20,] #aleatoriamente 20.

dato

## 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

## 7 4.6 3.4 1.4 0.3 setosa

## 8 5.0 3.4 1.5 0.2 setosa

## 9 4.4 2.9 1.4 0.2 setosa

## 10 4.9 3.1 1.5 0.1 setosa

## 11 5.4 3.7 1.5 0.2 setosa

## 12 4.8 3.4 1.6 0.2 setosa

## 13 4.8 3.0 1.4 0.1 setosa

## 14 4.3 3.0 1.1 0.1 setosa

## 15 5.8 4.0 1.2 0.2 setosa

## 16 5.7 4.4 1.5 0.4 setosa

## 17 5.4 3.9 1.3 0.4 setosa

## 18 5.1 3.5 1.4 0.3 setosa

## 19 5.7 3.8 1.7 0.3 setosa

## 20 5.1 3.8 1.5 0.3 setosa

Realizamos el gráfico:

**stem**(dato$Sepal.Length)

##

## The decimal point is at the |

##

## 4 | 34

## 4 | 6678899

## 5 | 00111444

## 5 | 778

**Gráficos de Dispersión**

**Matriz de Diagramas de Dispersión**

Para crear una matriz de diagramas de dispersión, se usa la siguiente sintaxis: pairs(x,…).

Como ejemplo: Realizar el gráfico de dispersión de la data iris usando el largo y el ancho del sépalo.

Recordar lo siguiente: - *pch* & *bg*: Color y símbolo - *oma*: Modificar el margen exterior

color<-**c**(2,3,4)[**unclass**(iris$Species)]

**pairs**(iris[1:2],

main="Diagrama de Dispersión",

pch=22, bg=color,

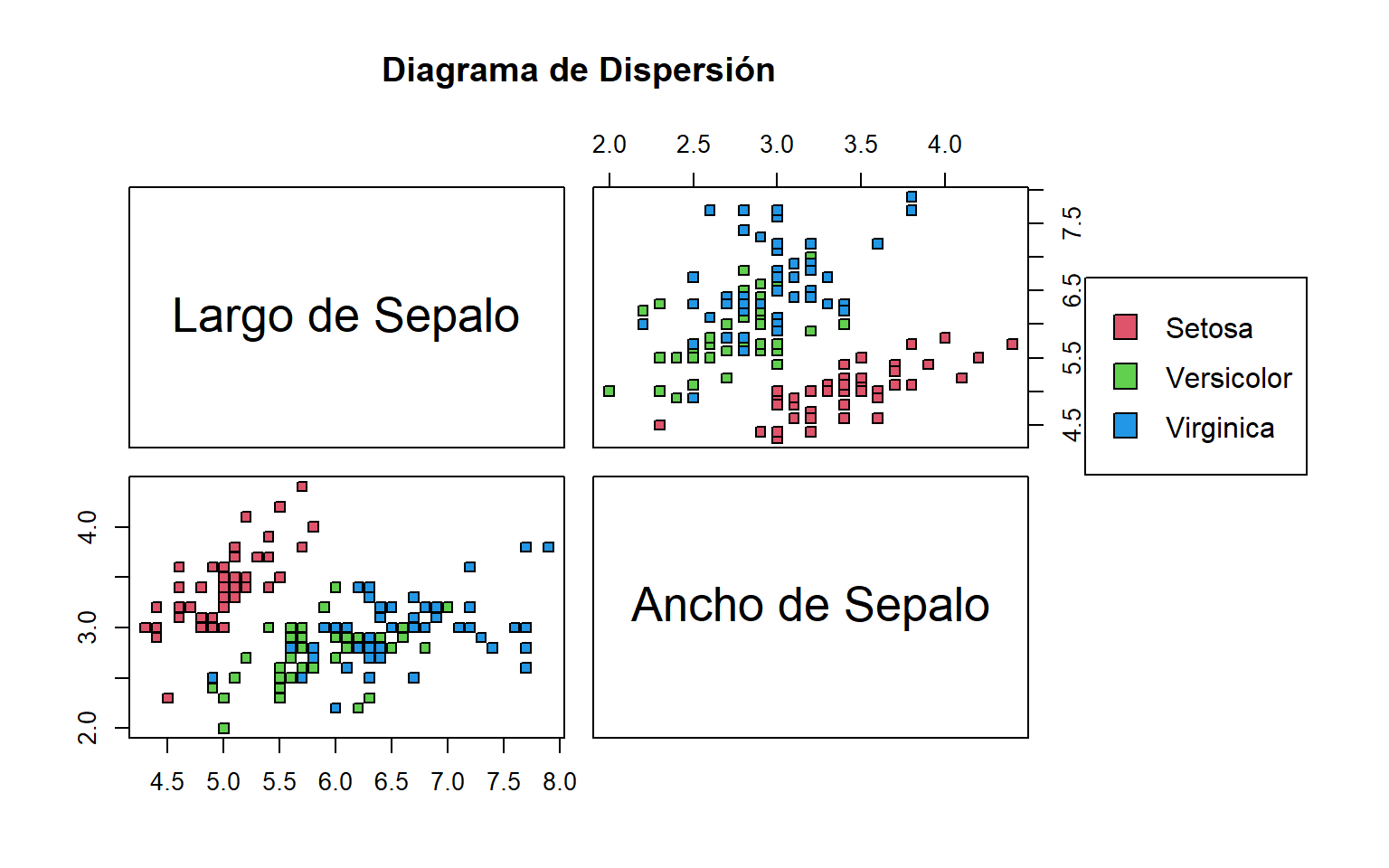
labels=c("Largo de Sepalo","Ancho de Sepalo"),

oma=c(4,4,6,12)

)

par(xpd=TRUE)

legend(0.8,0.7,c("Setosa","Versicolor","Virginica"), fill=c(2,3,4))

 Según el diagrama de dispersión, se puede observar que el ancho y largo del sépalo para la iris setosa se encuentra una correlación positiva y alta; mientras que en las otras dos variables, presentan una correlación moderada y positiva.

**Ejemplos Aplicados**

**Ejemplo 1: Diagrama de tallos y hojas**

Con los siguientes datos: 60 85 72 59 37 75 93 7 98 63 41 90 5 17 97. Realizar lo siguiente:

* Guardar los datos en archivo *txt* (bloc de notas)
* Importar dicho archivo en R.
* Obtener un resumen numérico de los datos.
* Hacer un diagrama de tallo y hojas.
* Elaborar una representación gráfica.

El archivo en el bloc de notas tiene que estar de la siguiente forma:

| **-** |
| --- |
| 60 |
| 85 |
| 72 |
| 59 |
| 37 |
| 75 |
| 93 |
| 7 |
| 98 |
| 63 |
| 41 |
| 90 |
| 5 |
| 17 |
| 97 |

Importamos los datos y posterior a ello vemos los estadísticos descriptivos más conocidos con *summary*:

data

## # A tibble: 15 x 1

## X1

## <dbl>

## 1 60

## 2 85

## 3 72

## 4 59

## 5 37

## 6 75

## 7 93

## 8 7

## 9 98

## 10 63

## 11 41

## 12 90

## 13 5

## 14 17

## 15 97

**summary**(data)

## X1

## Min. : 5.00

## 1st Qu.:39.00

## Median :63.00

## Mean :59.93

## 3rd Qu.:87.50

## Max. :98.00

Se realiza el diagrama de tallos y hojas

**stem**(data$X1,scale=2)

##

## The decimal point is 1 digit(s) to the right of the |

##

## 0 | 57

## 1 | 7

## 2 |

## 3 | 7

## 4 | 1

## 5 | 9

## 6 | 03

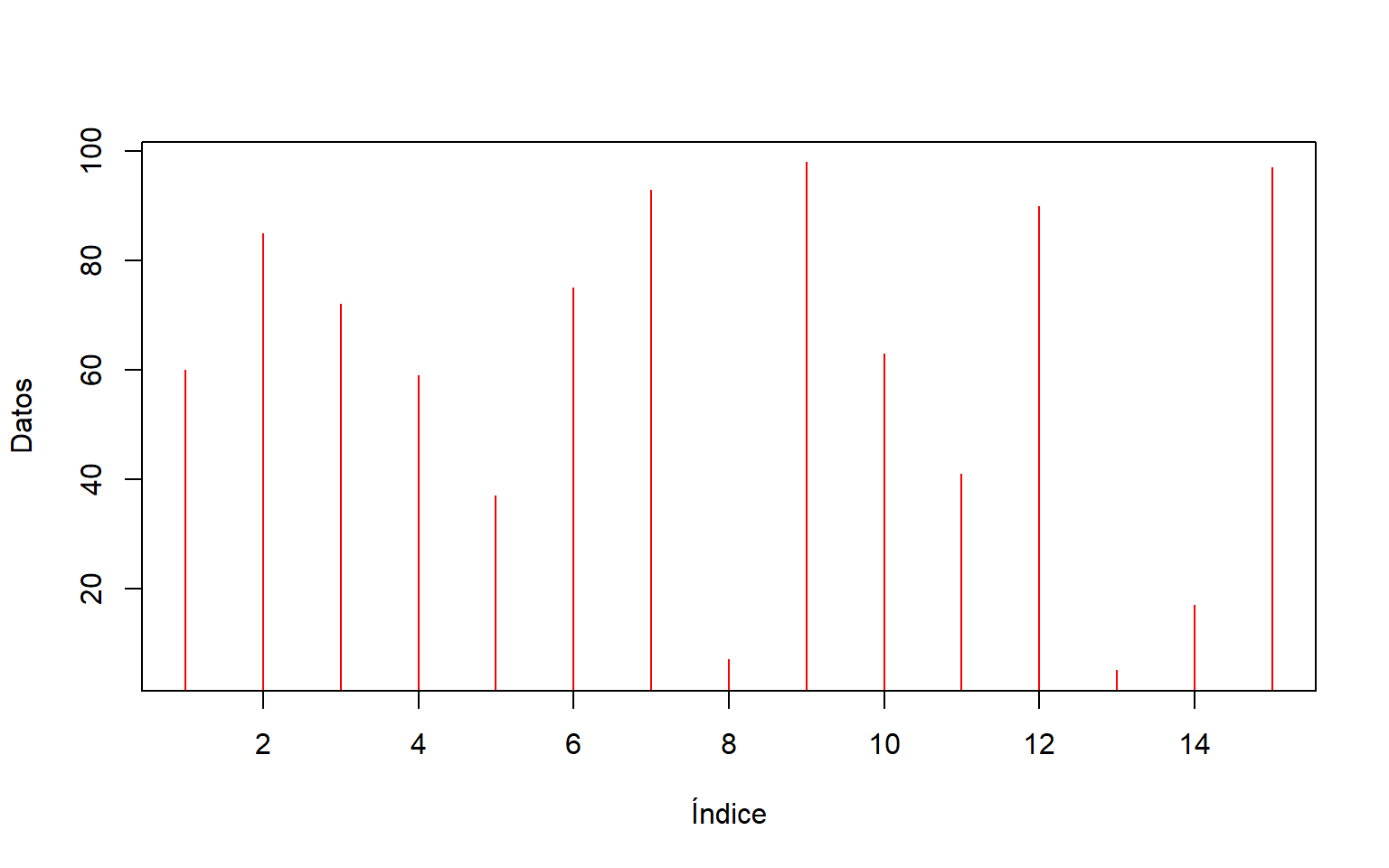
## 7 | 25

## 8 | 5

## 9 | 0378

La representación gráfica, como la data es cuantitativa y al presentar más de 8 grupos se podría hacer un diagrama de bastones.

**plot**(data$X1,type="h",col="red",xlab = "Índice",ylab = "Datos")



**Ejemplo 2: Gráfico de Barras**

Cargamos una base de datos producto que tiene los siguientes datos: los cuales son cualitativos.

Productos

## # A tibble: 50 x 1

## X1

## <chr>

## 1 A

## 2 B

## 3 A

## 4 A

## 5 C

## 6 B

## 7 C

## 8 C

## 9 B

## 10 C

## # ... with 40 more rows

Para relizar el gráfico de barras primero se debe hacer la tabal de frecuencias y finalmente con *barplot* se hace el gráfico de barras con las siguientes características.

tabla<-**table**(Productos)

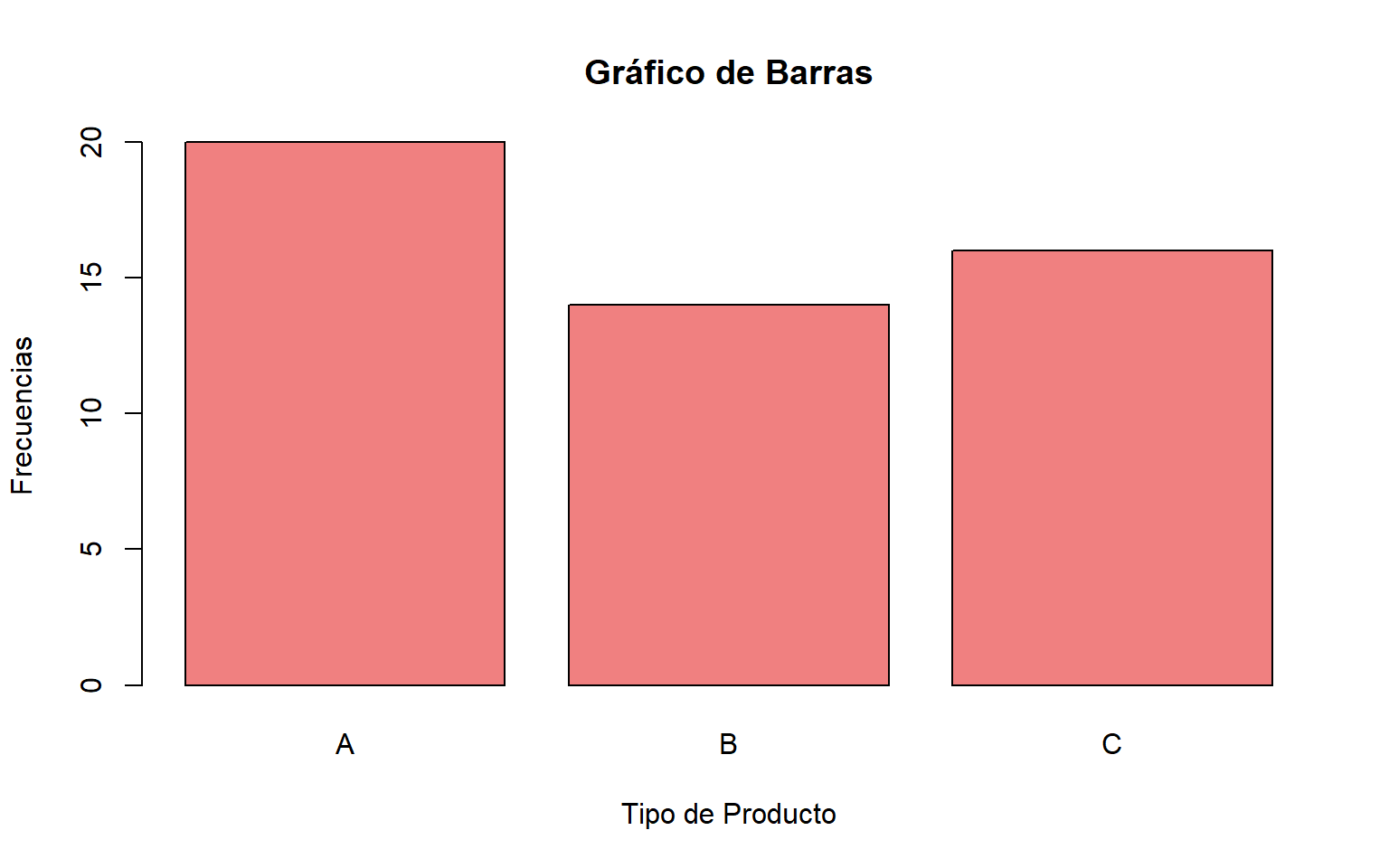
tabla

## Productos

## A B C

## 20 14 16

**barplot**(tabla, col ="lightcoral", main="Gráfico de Barras", xlab="Tipo de Producto",ylab="Frecuencias")



**Ejemplo 3: Diagrama Circular**

Considerar el ingreso de 200 trabajadores y con ello elaborar un gráfico circular donde se consideren grupos entre distintos rangos de sueldo.

Los datos son los siguientes:

ingreso

## # A tibble: 200 x 1

## X1

## <dbl>

## 1 1000

## 2 1400

## 3 1350

## 4 930

## 5 930

## 6 1200

## 7 1050

## 8 930

## 9 953

## 10 974

## # ... with 190 more rows

Creamos la variable ingreso en referencia a cada grupo:

dato<-ingreso #Llamamos a los ingresos como "dato"

dato$ingreso<-ingreso$X1 #La columna de ingreso X1 la llamaremos "dato$ingreso"

ingreso<-**rep**(NA, **length**(dato))

ingreso[dato==930]<-'=930'

ingreso[dato>=931 & dato<=1500]<-'entre 931 y 1500'

ingreso[dato>=1501 & dato<=3000]<-'entre 1501 y 3000'

ingreso[dato>=3001 & dato<=5000]<-'entre 3001 y 5000'

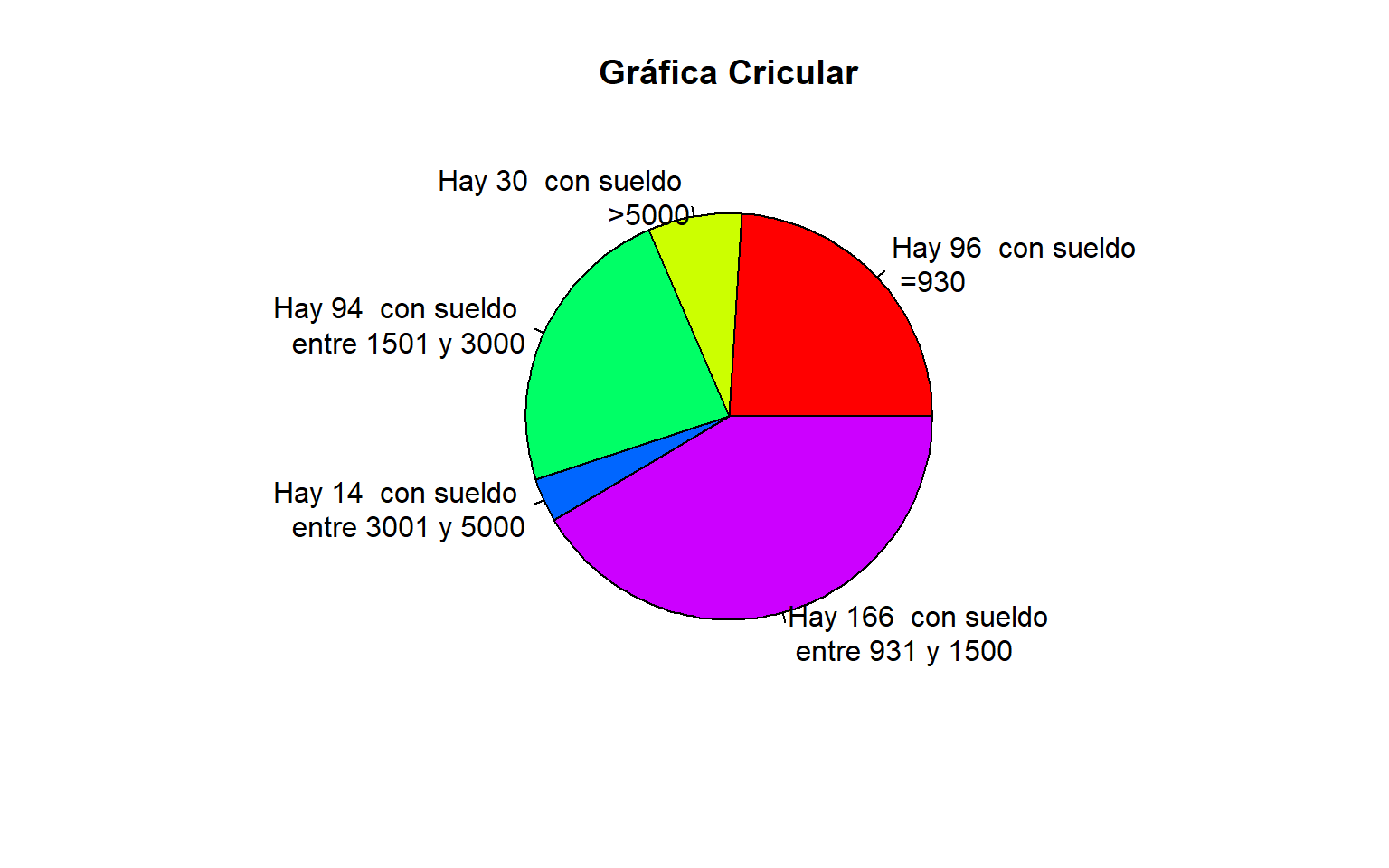
ingreso[dato>=5001]<-'>5000'

Calculamos las frecuencias y etiquetas:

frecuencia<-**table**(ingreso)

etiquetas<-**paste**("Hay",frecuencia," con sueldo \n", **names**(frecuencia))

**pie**(frecuencia, main ="Gráfica Cricular", col=**rainbow**(**length**(frecuencia)), labels=etiquetas)



**Ejemplo 4: Diagrama de Cajas con GGPUBR**

Instalamos el paquete “ggpubr” y cargamos las librerías:

*install.package(“ggpubr”)*

*library(ggpubr)*

*library(ggplot2)*

*library(magrittr)*

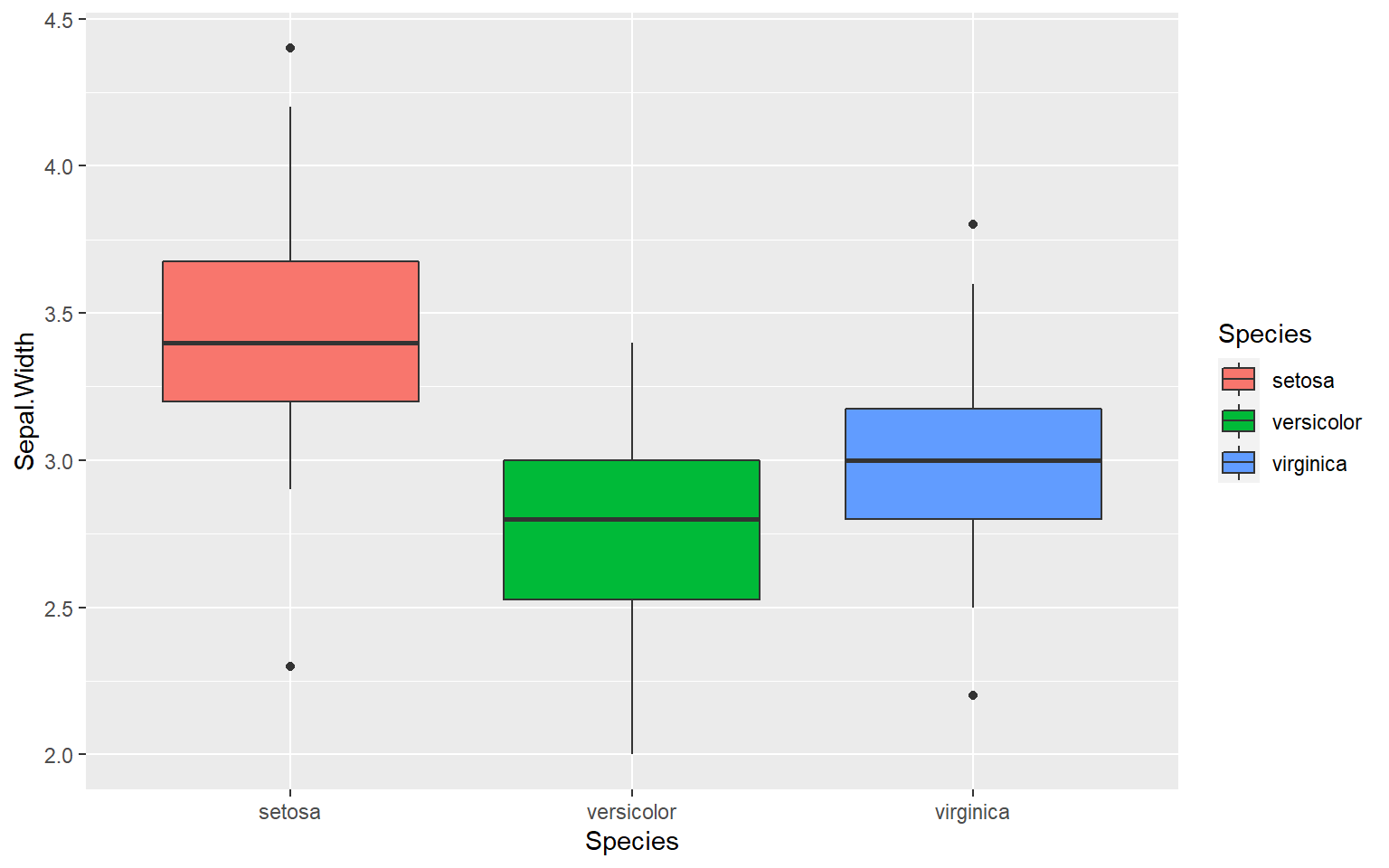
**library**(ggplot2)

**library**(magrittr)

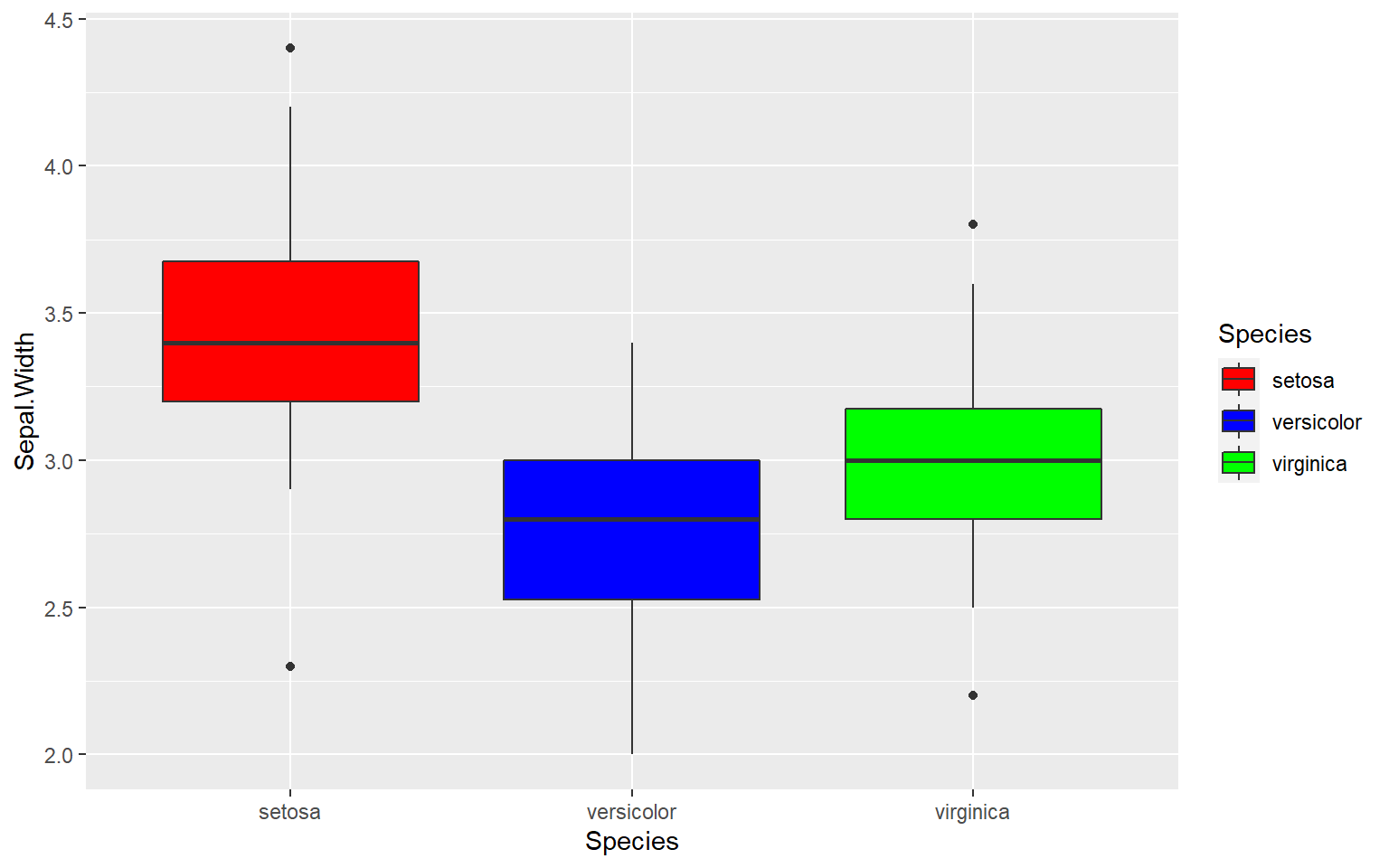
**library**(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.0.2

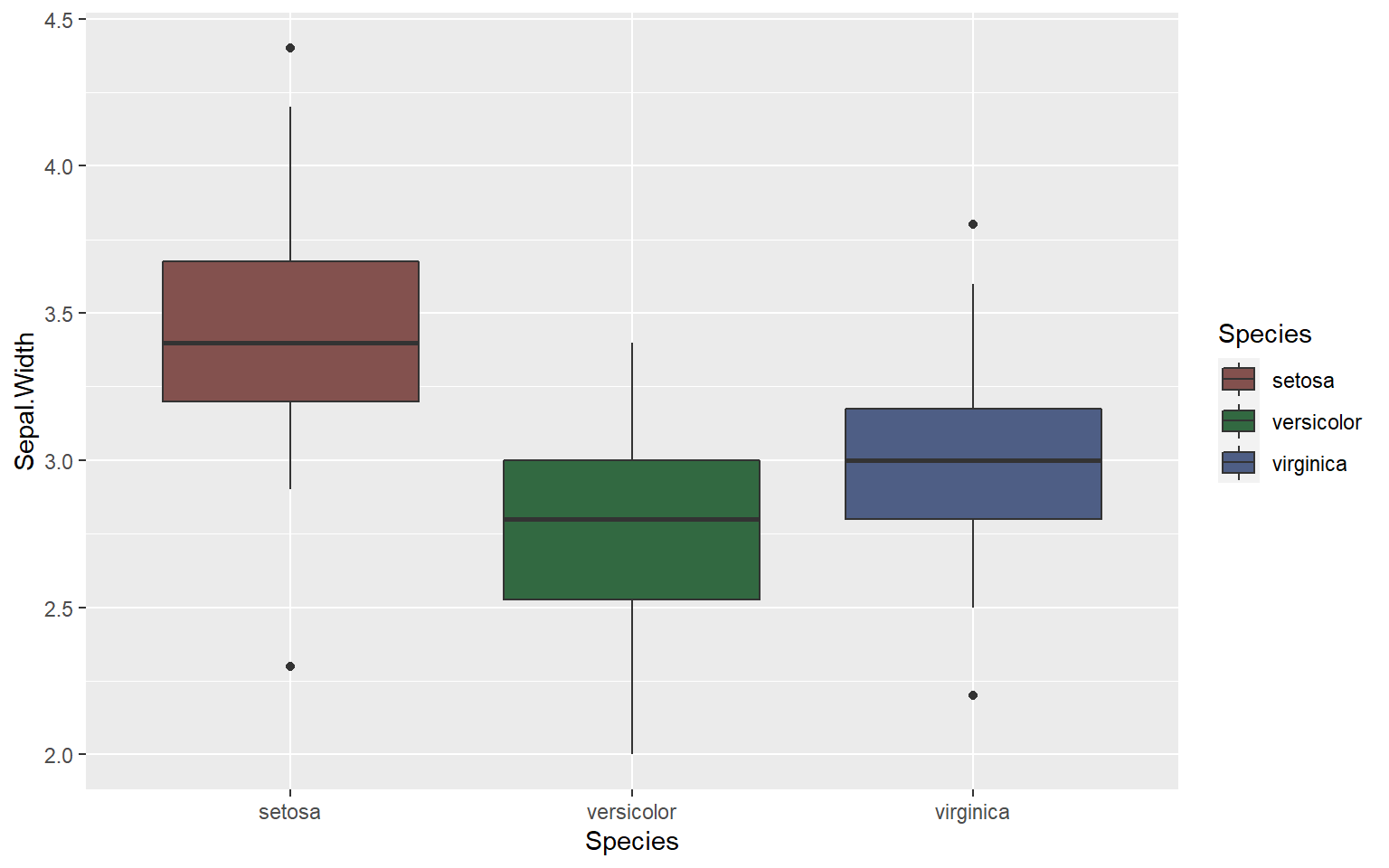
**ggplot**(iris,**aes**(x=Species,y=Sepal.Width, fill=Species))+**geom\_boxplot**()

 Modificamos el color de cada caja:

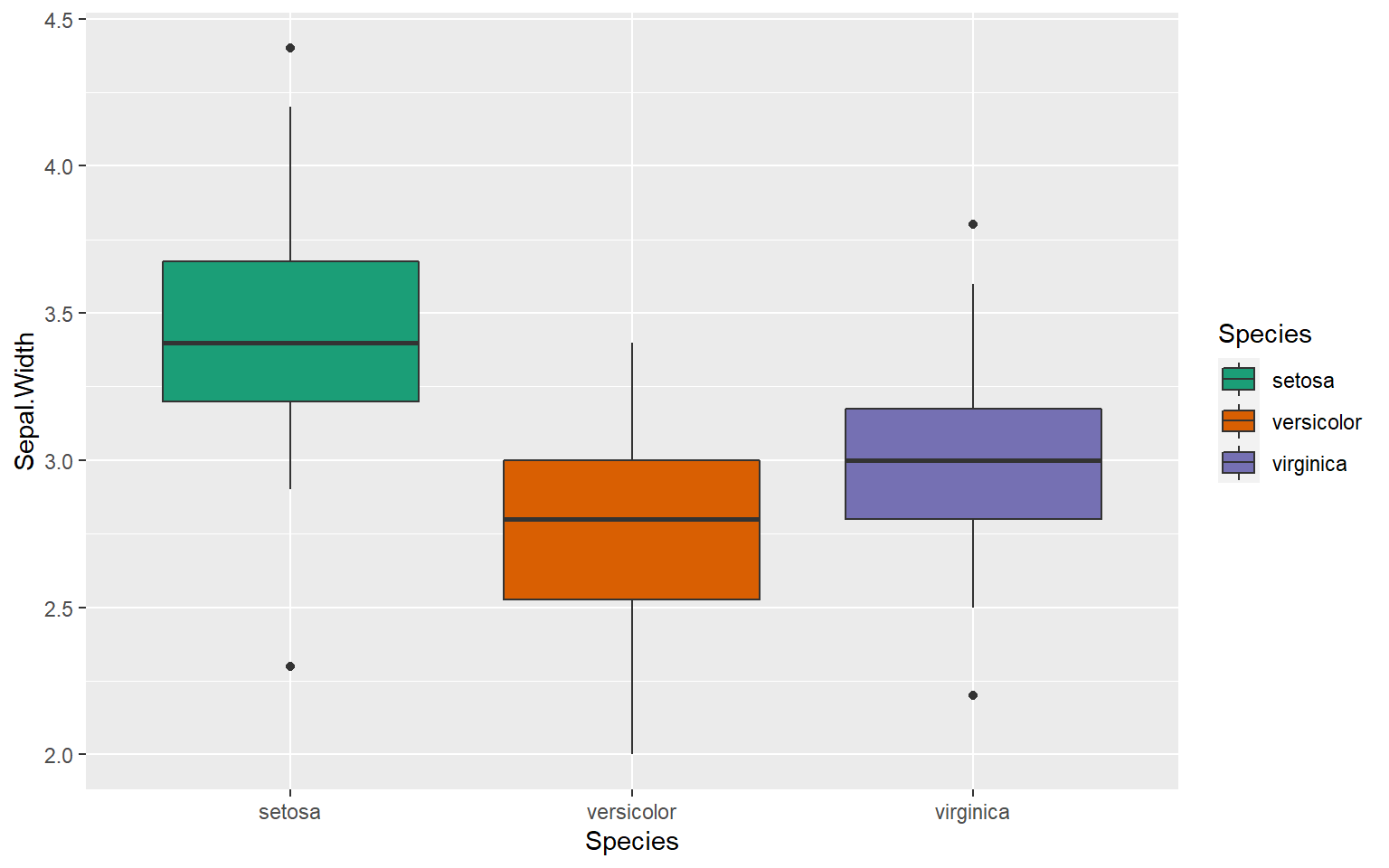
**ggplot**(iris, **aes**(x=Species, y=Sepal.Width, fill=Species)) + **geom\_boxplot**() + **scale\_fill\_manual**(values=**c**("red","blue","green"))

 Otro:

**ggplot**(iris, **aes**(x=Species, y=Sepal.Width, fill=Species)) + **geom\_boxplot**() + **scale\_fill\_hue**(l=40, c=35)

 Otro:

**ggplot**(iris, **aes**(x=Species, y=Sepal.Width, fill=Species)) + **geom\_boxplot**() + **scale\_fill\_brewer**(palette="Dark2")



**Ejemplo 5: Gráficos de Violín**

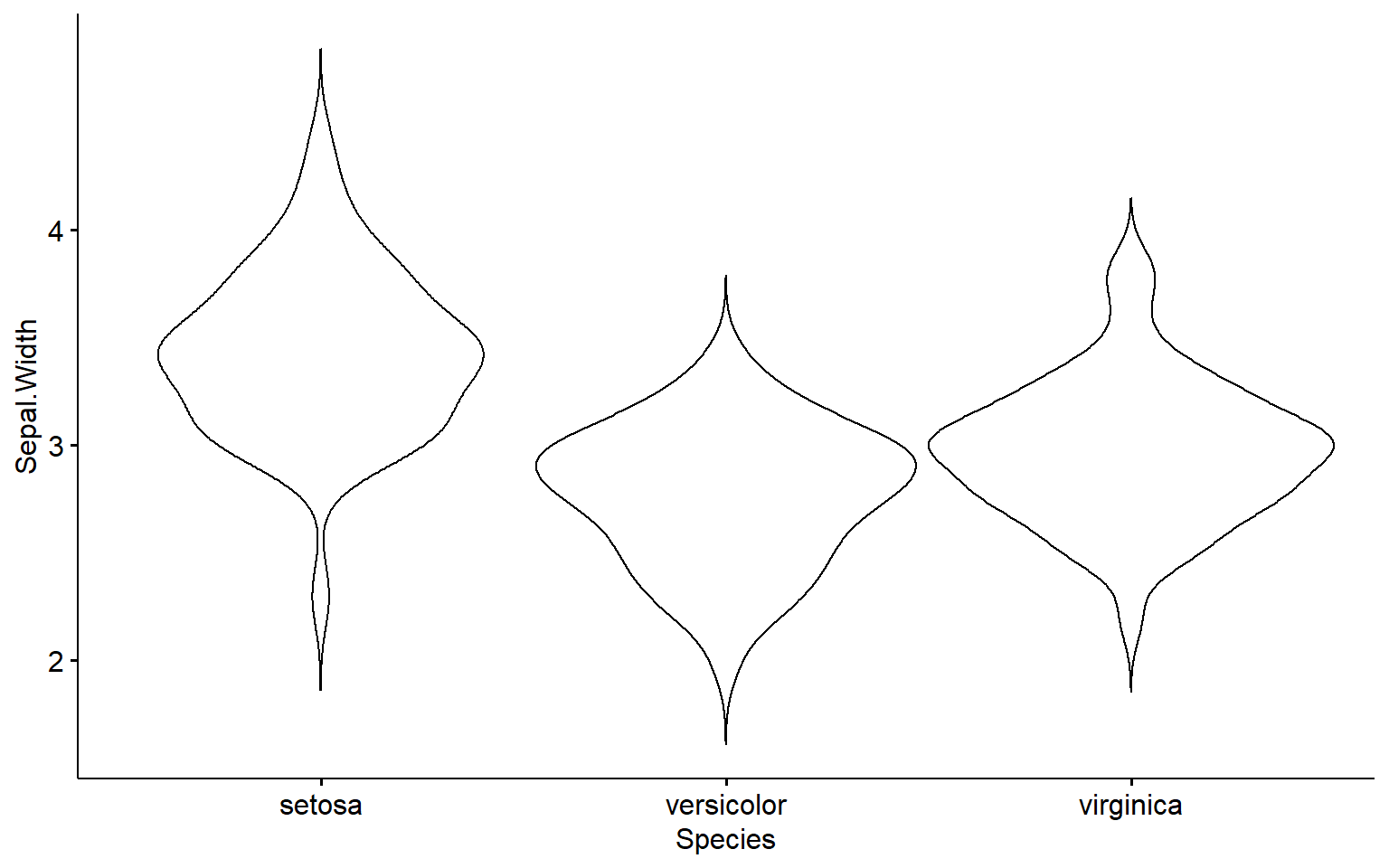
Para este ejemplo se trabajara con la dataset iris.

**library**(ggplot2)

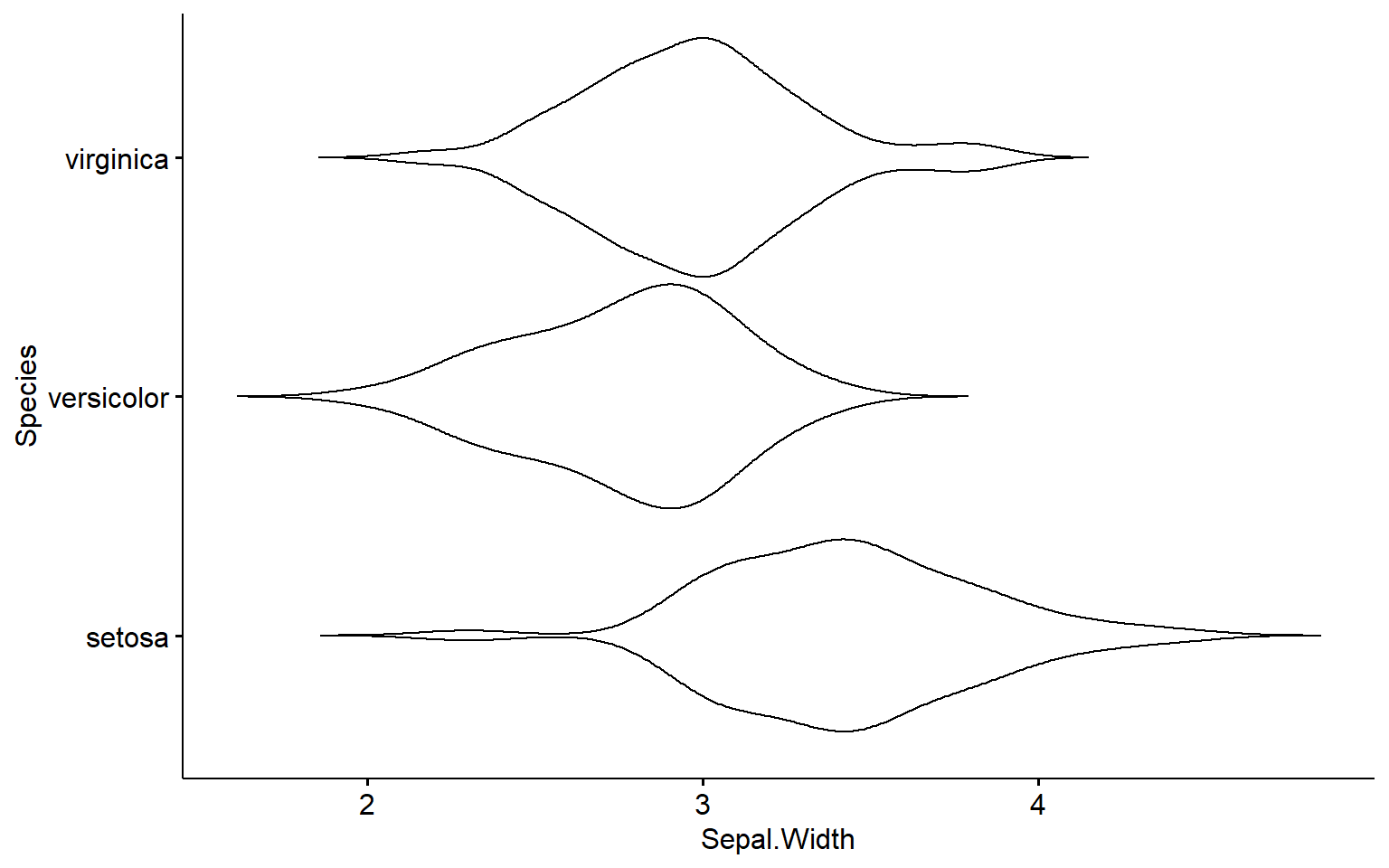
**library**(magrittr)

**library**(ggpubr)

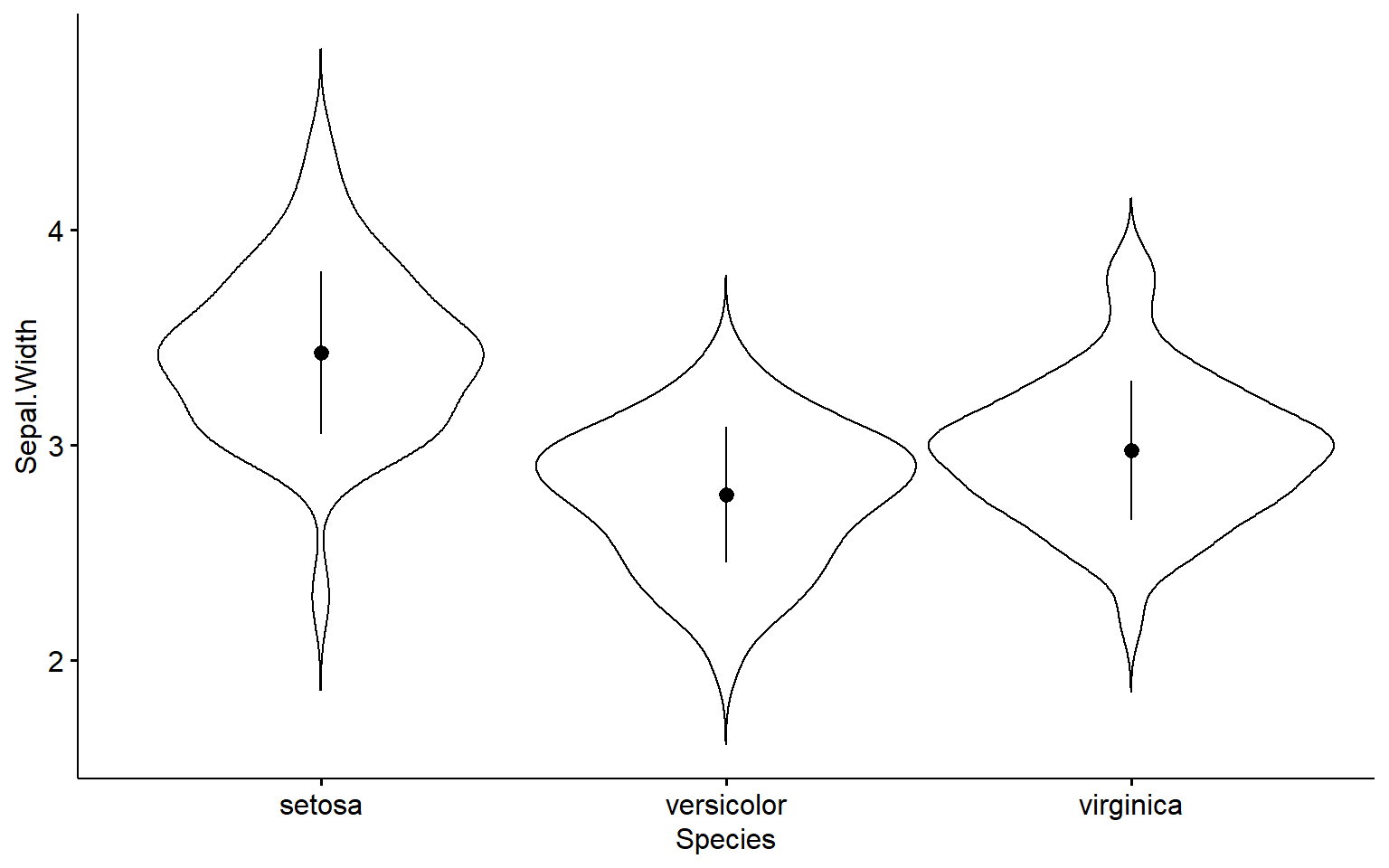
**ggviolin**(iris, x="Species",y="Sepal.Width")

 En forma horizontal:

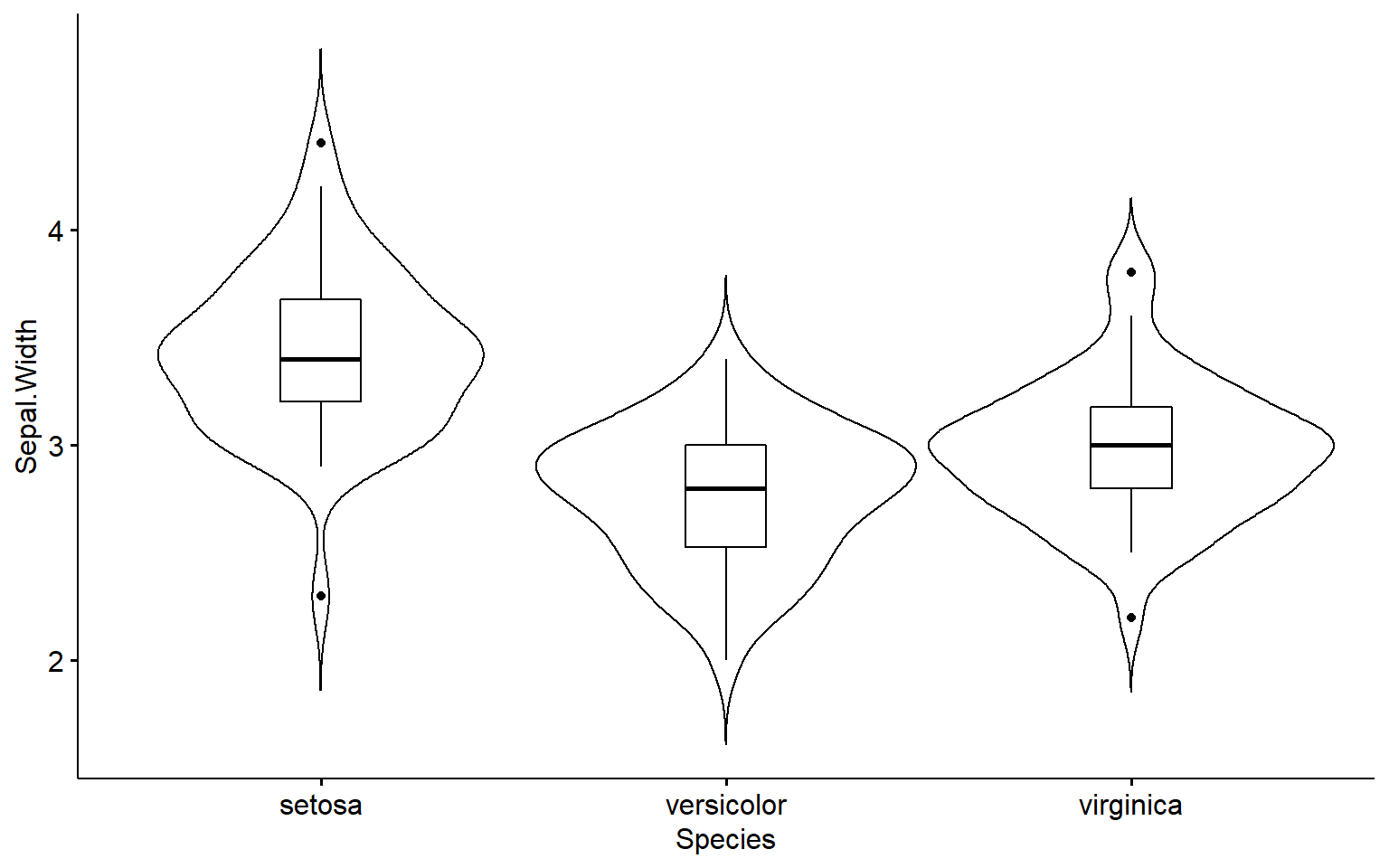
**ggviolin**(iris, x="Species",y="Sepal.Width", orientation="horiz")

 Agregando la desviación estándar:

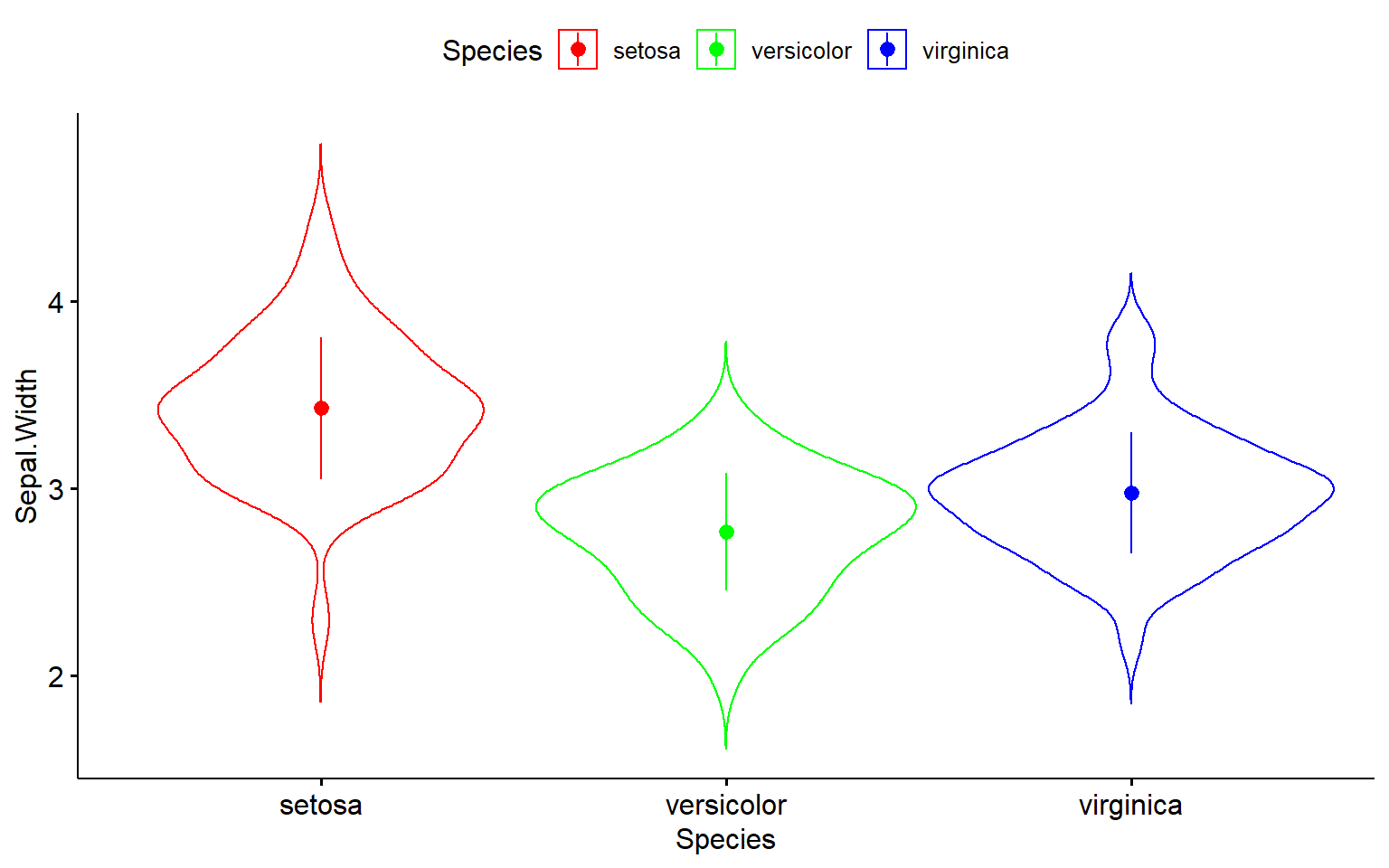
**ggviolin**(iris, x="Species",y="Sepal.Width", add="mean\_sd")

 Agregamos el bloxpot:

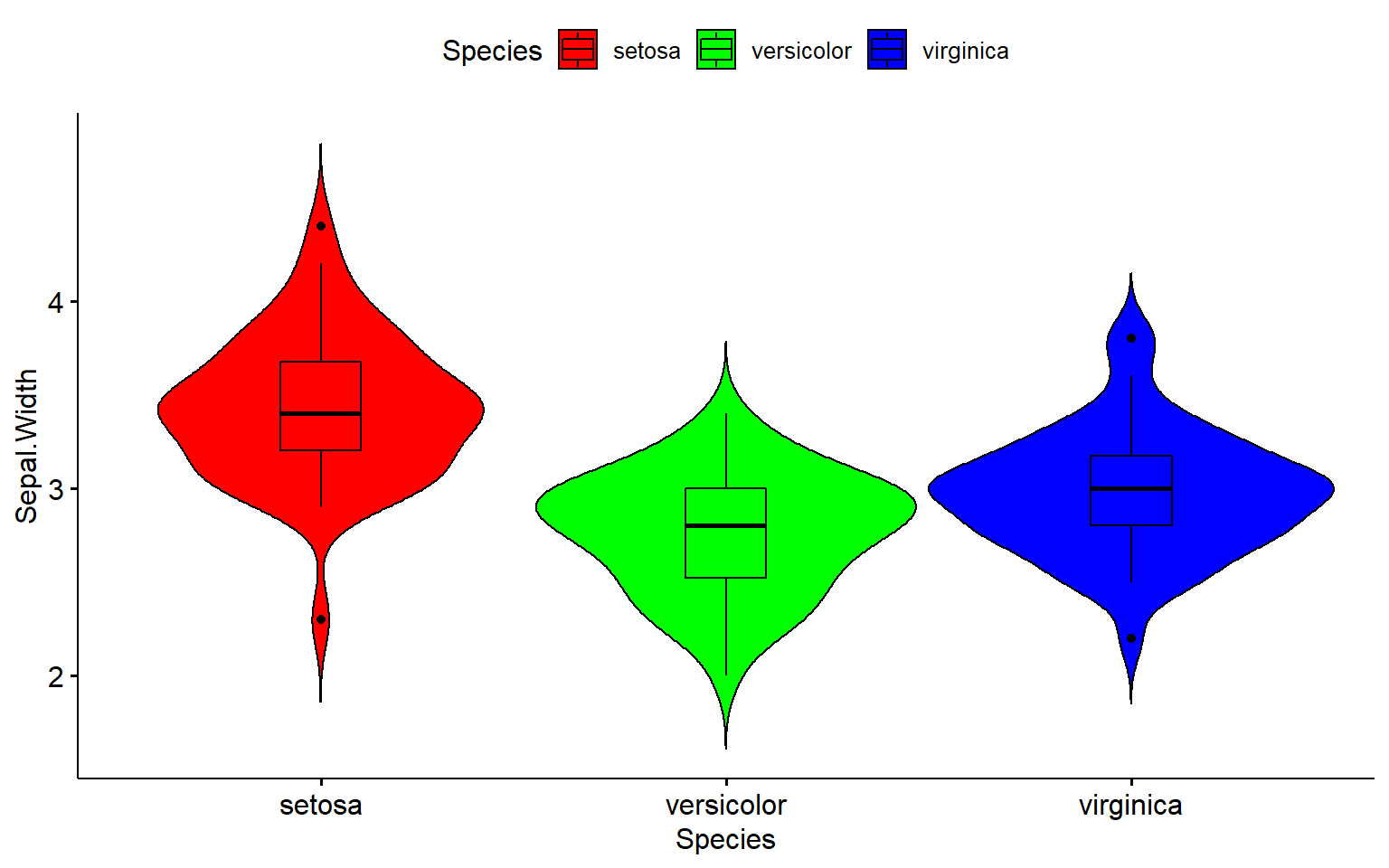
**ggviolin**(iris, x="Species",y="Sepal.Width", add="boxplot")

 Otro con colores:

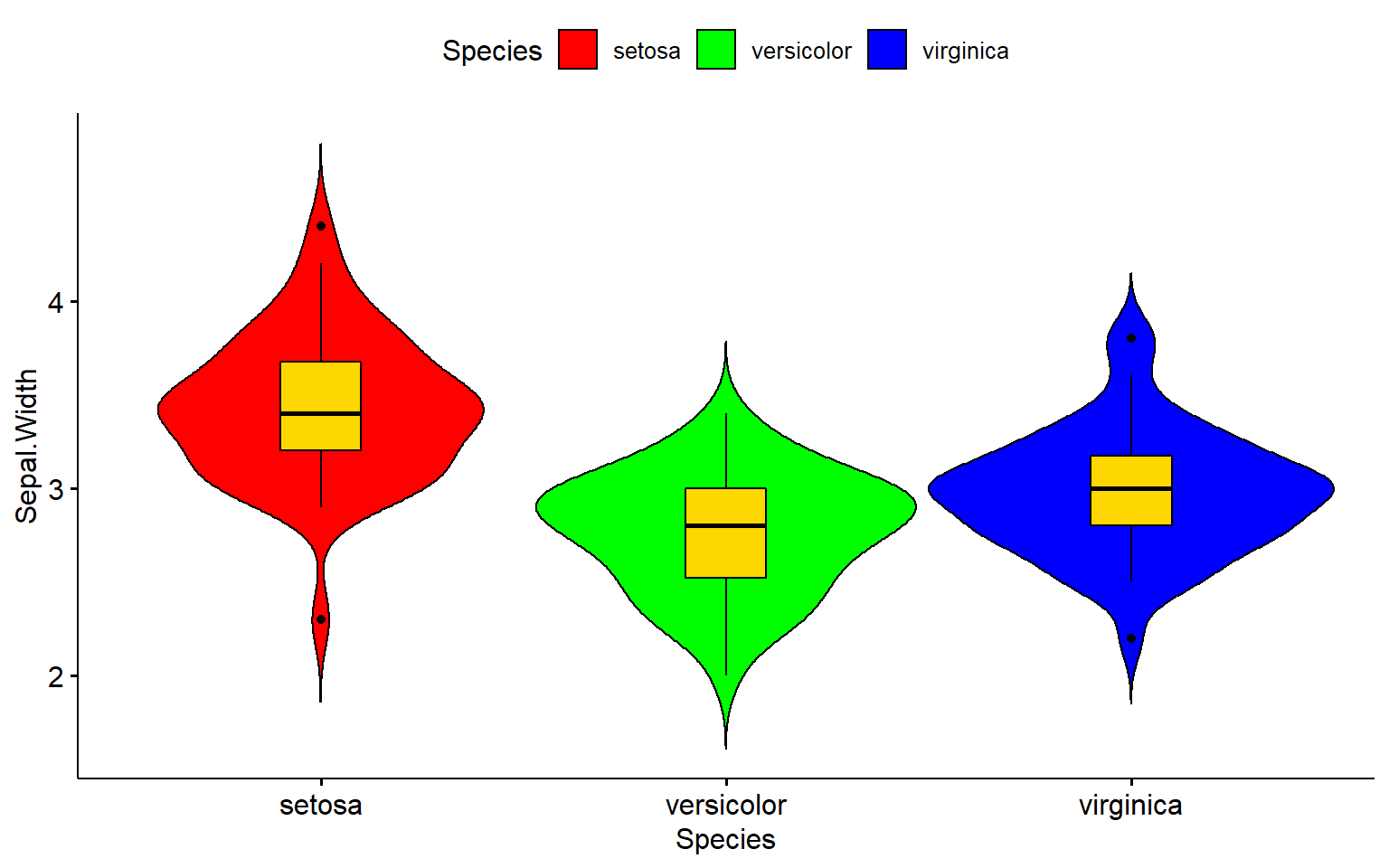
**ggviolin**(iris, x="Species",y="Sepal.Width",color ="Species",palette=**rainbow**(3),add="mean\_sd")

 Otro:

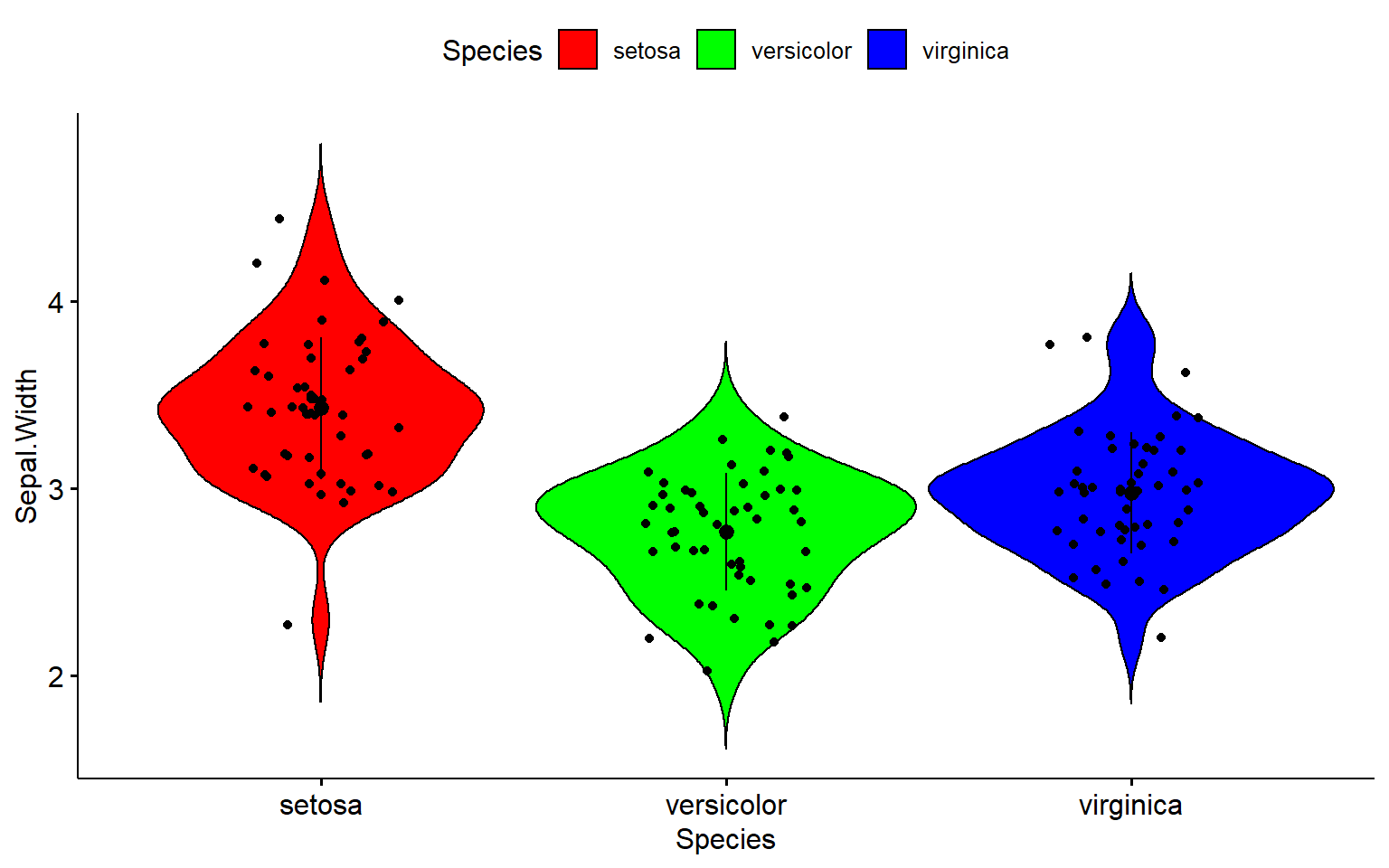
**ggviolin**(iris, x="Species",y="Sepal.Width",fill="Species",palette=**rainbow**(3),add="boxplot")

 Otro:

**ggviolin**(iris, x="Species",y="Sepal.Width",fill="Species", palette=**rainbow**(3), add ="boxplot", add.params=**list**(fill="gold"))

 Otro:

**ggviolin**(iris, x="Species",y="Sepal.Width",fill="Species",palette=**rainbow**(3),add=**c**("jitter","mean\_sd"))



**Ejemplo 6: Gráficos de Dispersión**

Primero importamos una data que tenga las siguientes características entre las edades y numero de hijos de 50 personas.

persona

## # A tibble: 50 x 2

## Edad Peso

## <dbl> <dbl>

## 1 20 0

## 2 20 0

## 3 20 0

## 4 20 1

## 5 20 1

## 6 20 0

## 7 20 0

## 8 23 0

## 9 23 1

## 10 23 0

## # ... with 40 more rows

Calculamos la frecuencia:

T<-**xyTable**(persona$Edad,persona$Peso)

T

## $x

## [1] 20 20 23 23 27 27 33 33 39 39 39 45 45 45 50 50 50 65 65

##

## $y

## [1] 0 1 0 1 2 3 2 3 2 3 4 2 3 4 3 4 5 4 5

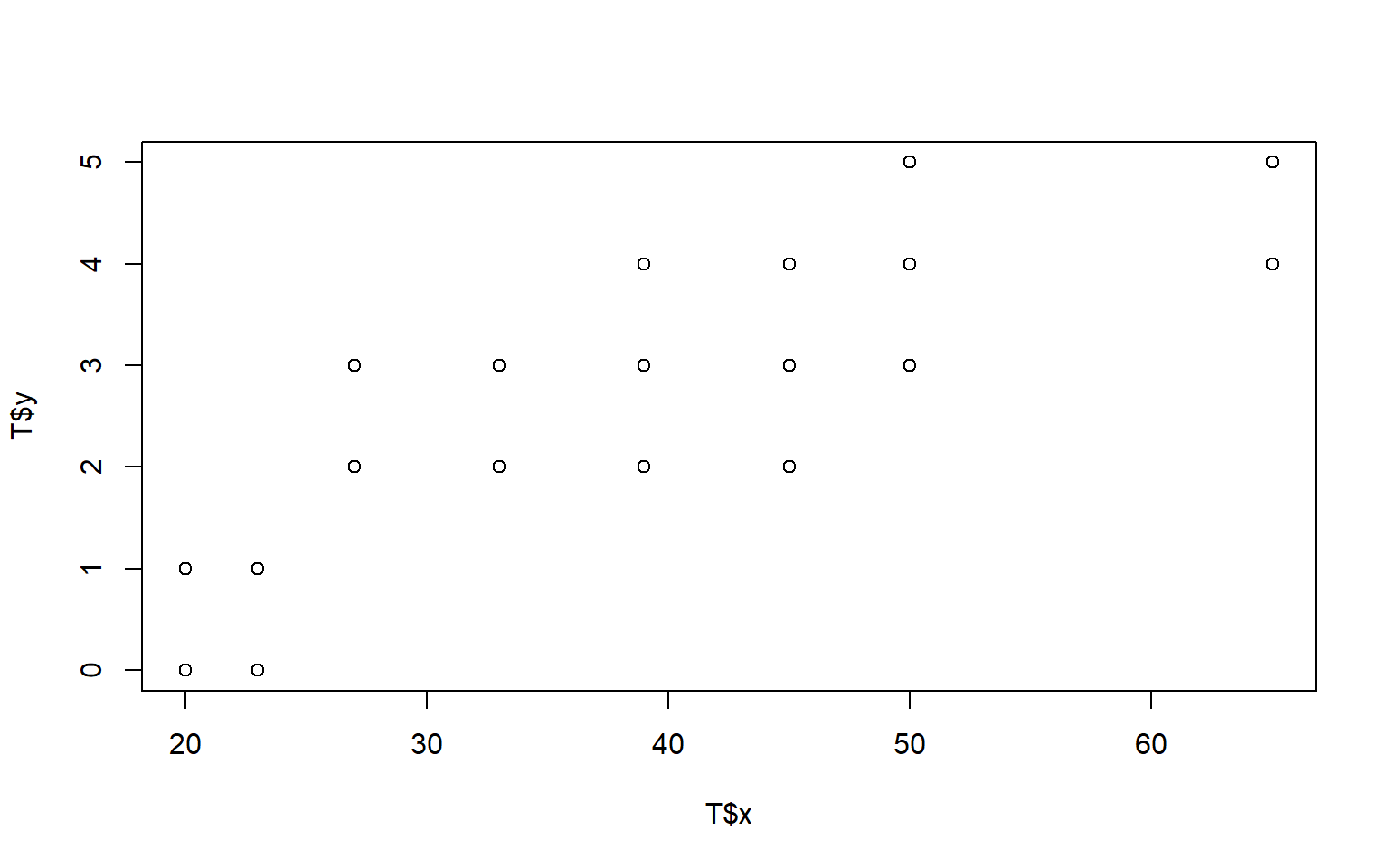
##

## $number

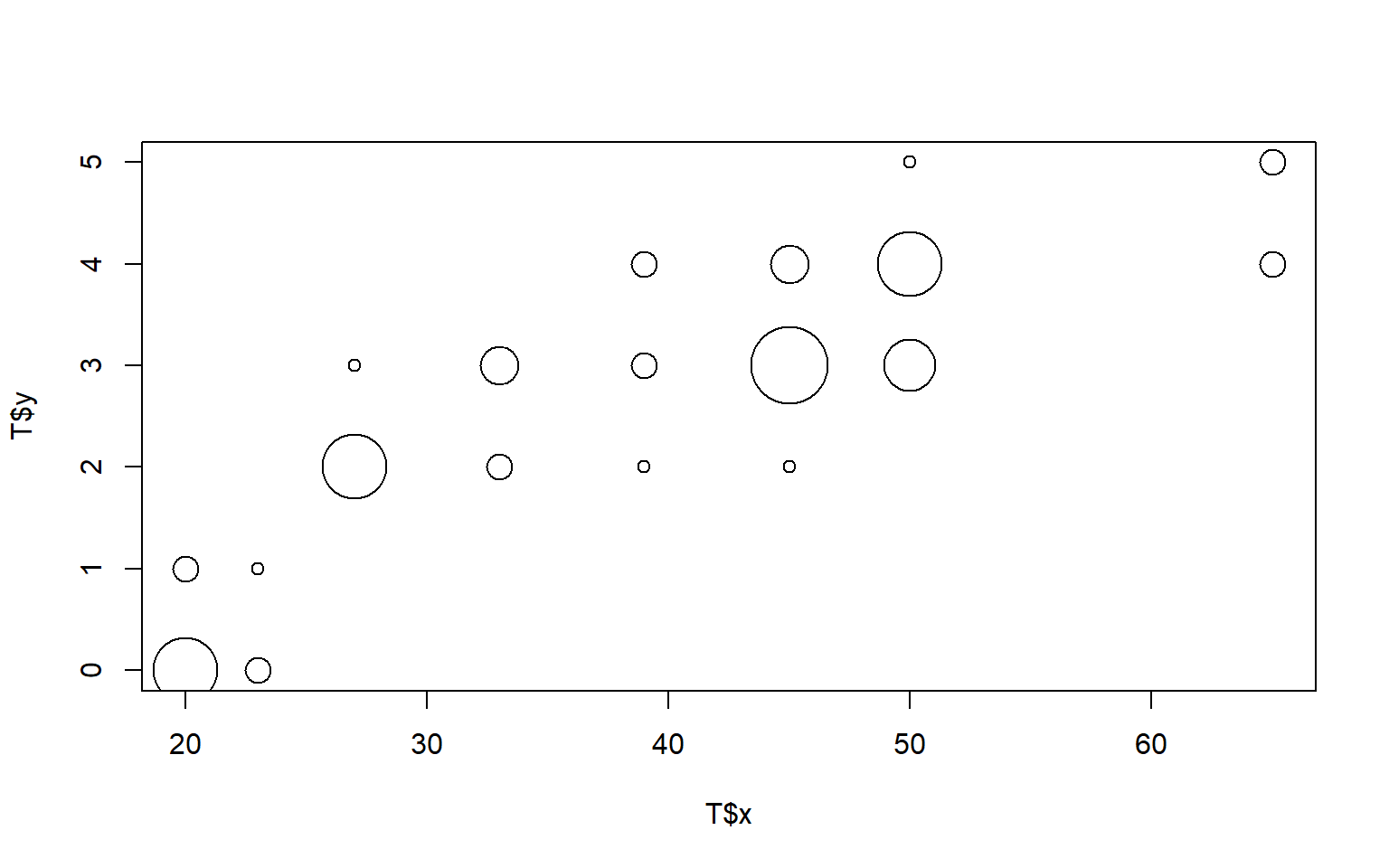
## [1] 5 2 2 1 5 1 2 3 1 2 2 1 6 3 4 5 1 2 2

La gráfica de dispersión por defecto:

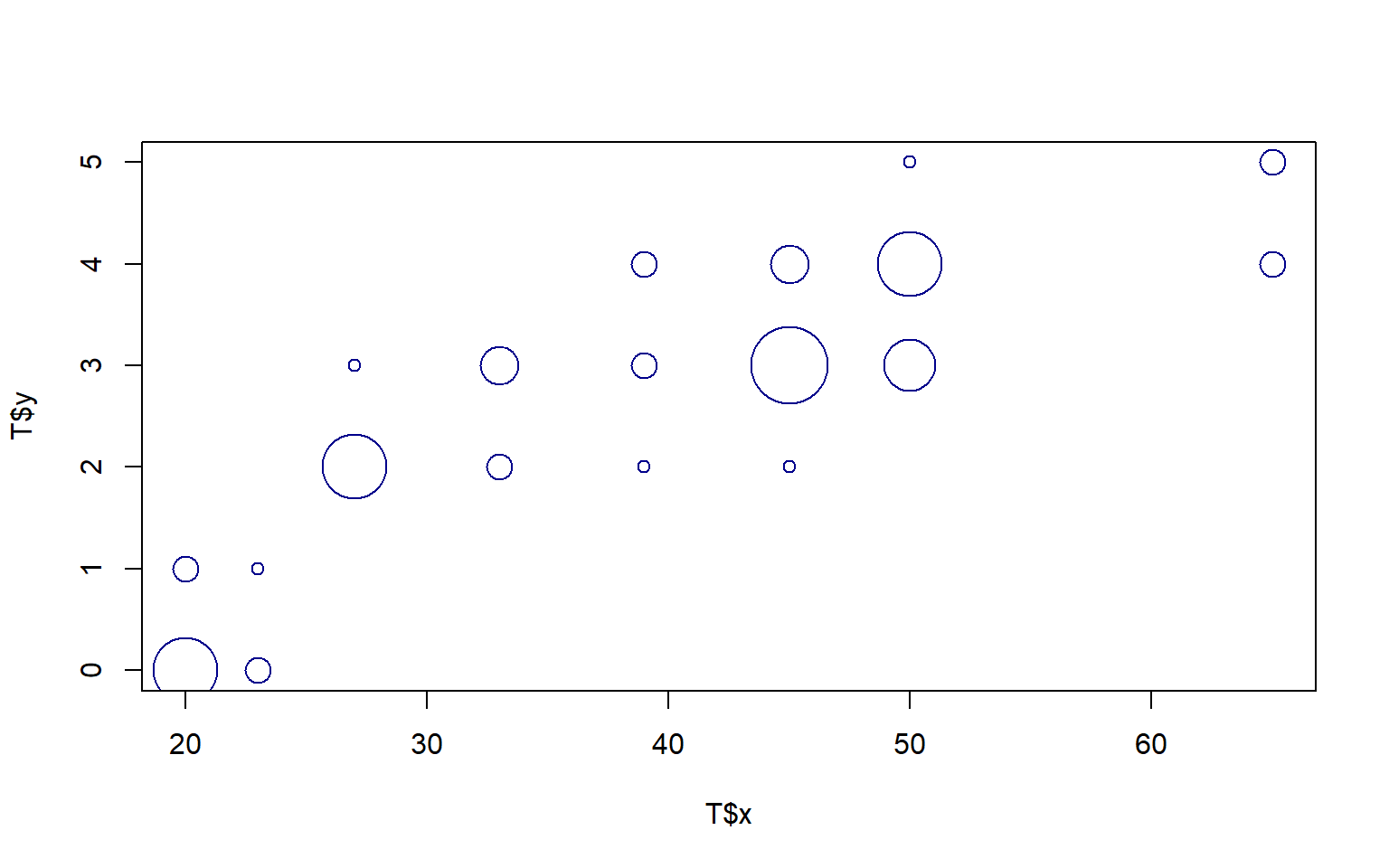
**plot**(T$x,T$y)

 Modificando el tamaño de los marcadores:

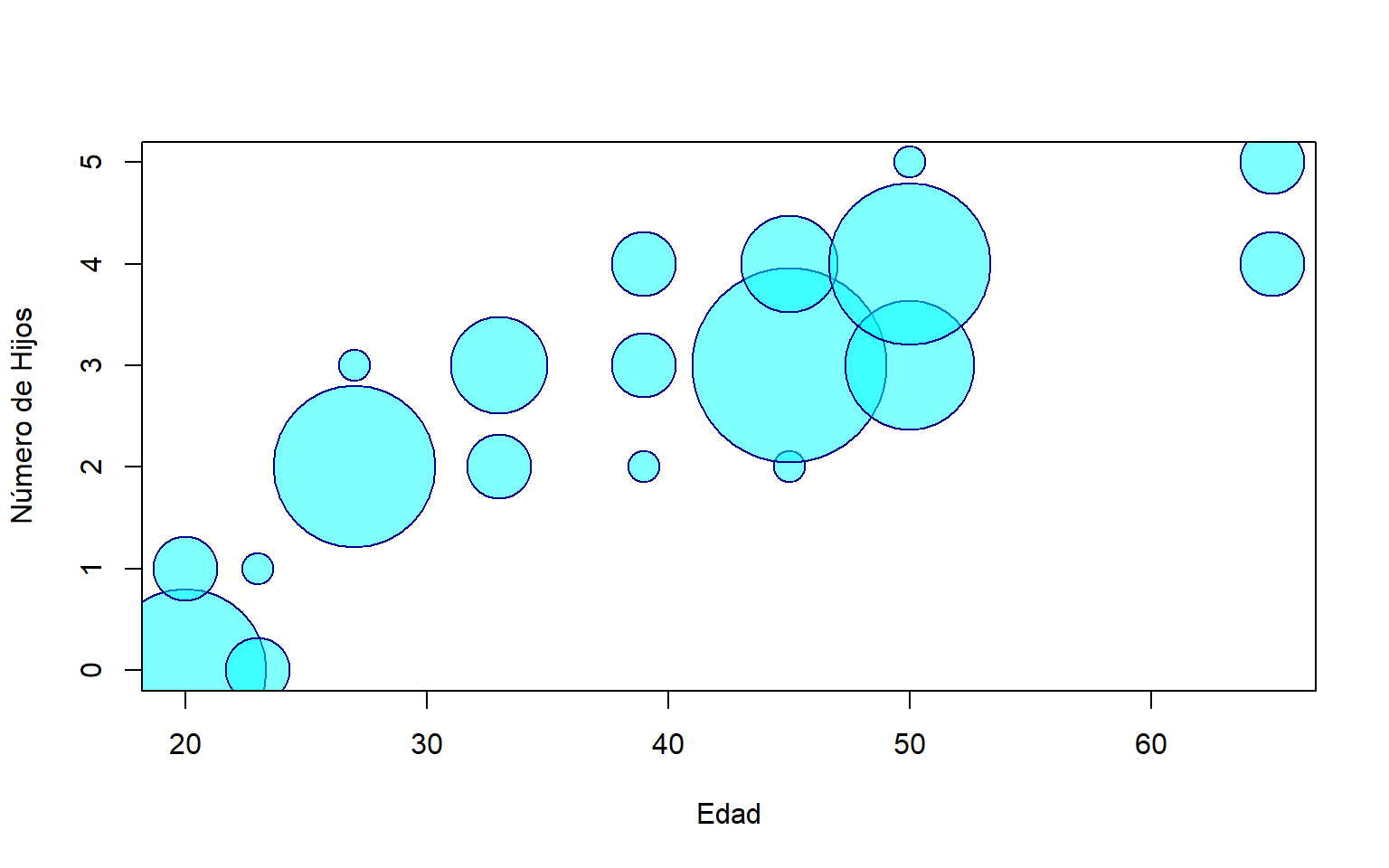
**plot**(T$x, T$y, cex=T$number)

 Modificando el color:

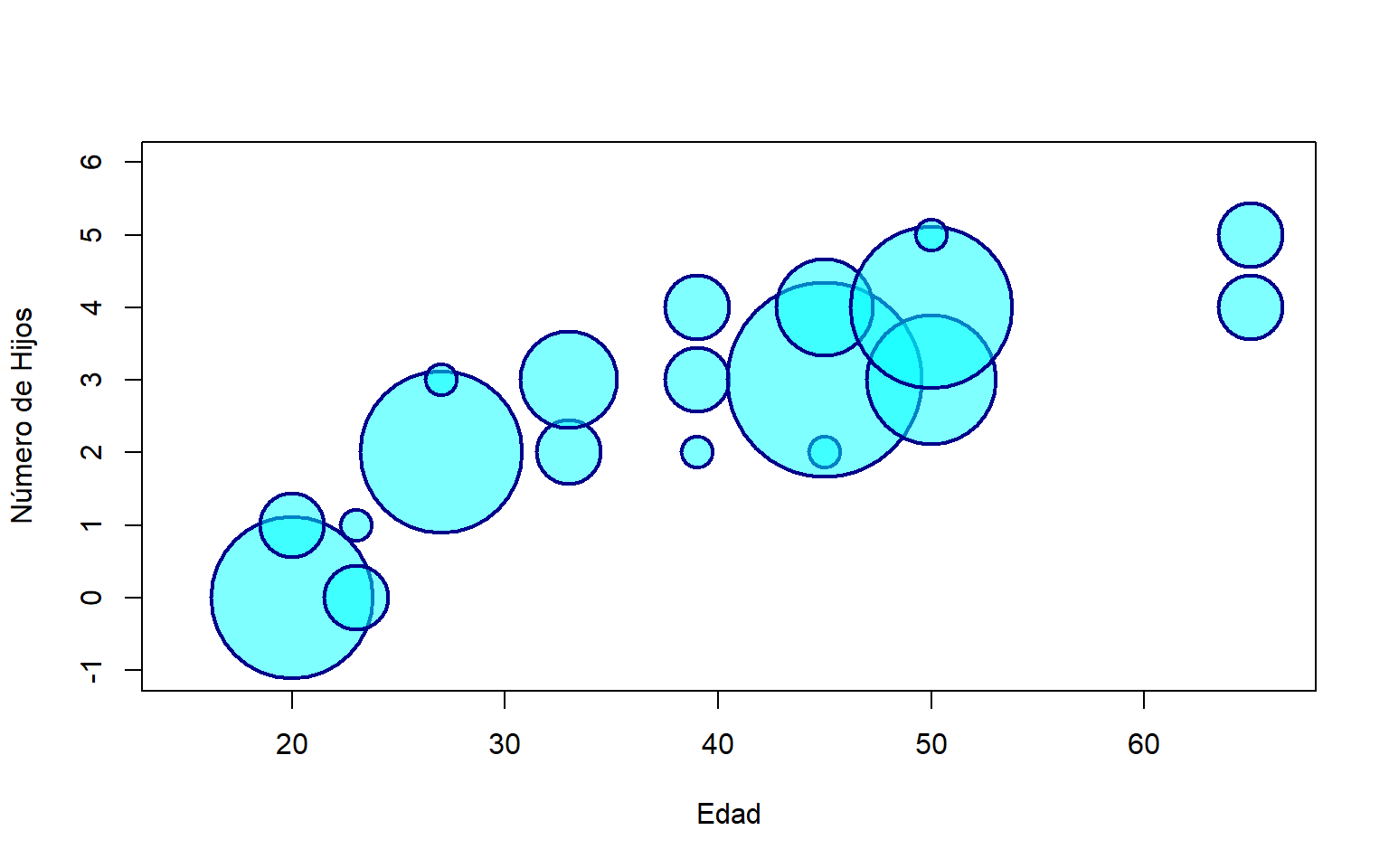
**plot**(T$x, T$y, cex=T$number, pch=21, col="darkblue")

 Otro:

**plot**(T$x, T$y, cex=2.5\*T$number, pch=21, col="darkblue", bg=**rgb**(0,1,1,0.5),xlab="Edad",ylab="Número de Hijos")

 Modificando el grosor de la línea:

**plot**(T$x, T$y, cex=2.5\*T$number, pch=21, col="darkblue", bg=**rgb**(0,1,1,0.5),lwd=2,xlab="Edad",ylab="Número de Hijos",xlim=**c**(15,66),ylim=**c**(-1,6))

 Ingresando Texto:

**plot**(T$x, T$y, cex=2.5\*T$number, pch=21, col="darkblue", bg=**rgb**(0,1,1,0.5),lwd=2,xlab="Edad",ylab="Número de Hijos",xlim=**c**(15,66),ylim=**c**(-1,6))

**text**(T$x, T$y, T$number)

