EDA_R

Andrea B

2023-11-14

Gráfico de Densidad

Un grafico de densidad es una representación de la distribución de una variable numérica.Es una versión suavizada del histograma y se utiliza en el mismo tipo de situación. A continuación se muestra un ejemplo básico creado con la biblioteca.

```
# Libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

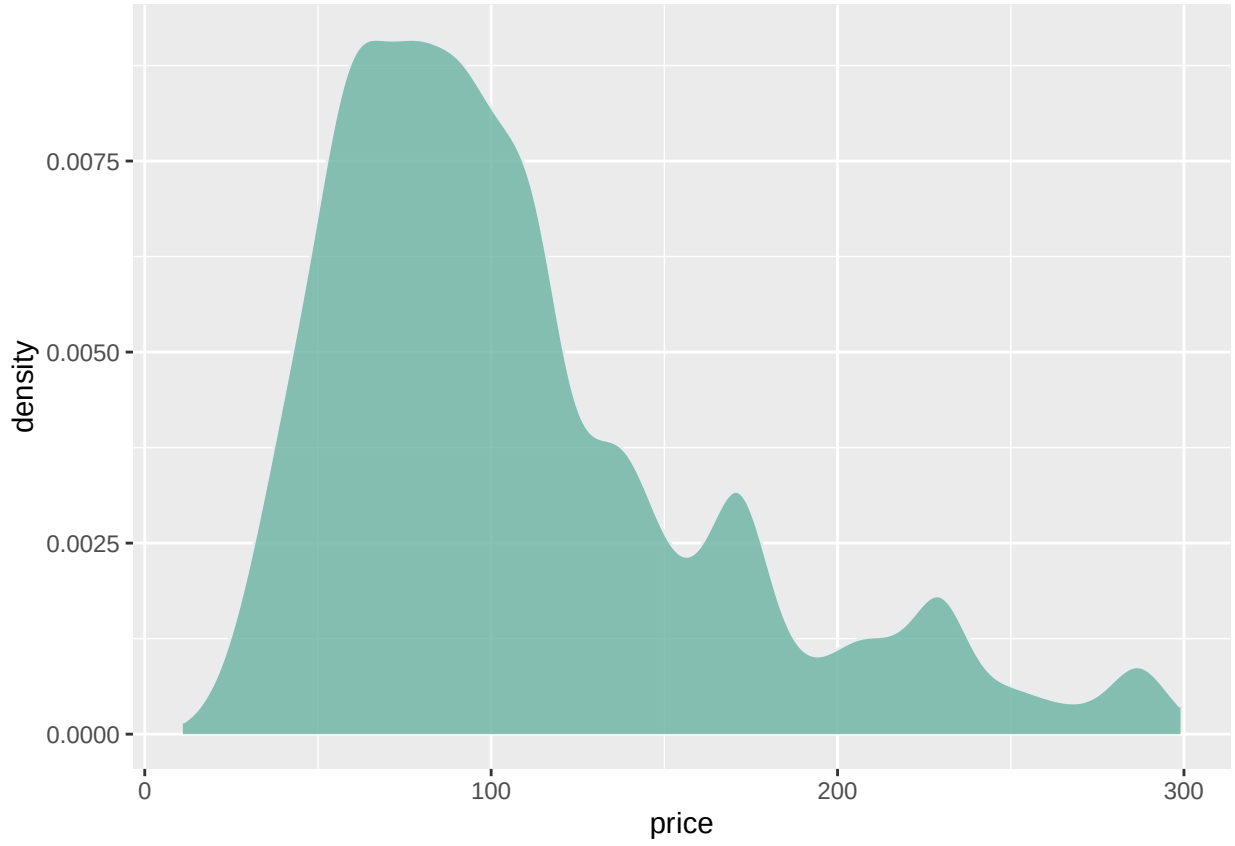
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load dataset from github
data <- read.table("https://raw.githubusercontent.com/holtzy/data_to_viz/master/Example_dataset/1_OneNu

# Make the histogram
data %>%
  filter( price<300 ) %>%
  ggplot( aes(x=price)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)
```



Histograma con Varios grupos

Se utilizan comúnmente en el análisis de datos para observar la distribución de variables. Una tarea común en la visualización de datos es comparar la distribución de 2 variables simultáneamente.

```
#Create data
set.seed(1)
Ixos=rnorm(4000 , 120 , 30)
Primadur=rnorm(4000 , 200 , 30)

# First distribution
hist(Ixos, breaks=30, xlim=c(0,300), col=rgb(1,0,0,0.5), xlab="height",
     ylab="nbr of plants", main="distribution of height of 2 durum wheat varieties" )

# Second with add=T to plot on top
hist(Primadur, breaks=30, xlim=c(0,300), col=rgb(0,0,1,0.5), add=T)

# Add legend
legend("topright", legend=c("Ixos","Primadur"), col=c(rgb(1,0,0,0.5),
     rgb(0,0,1,0.5)), pt.cex=2, pch=15 )
```

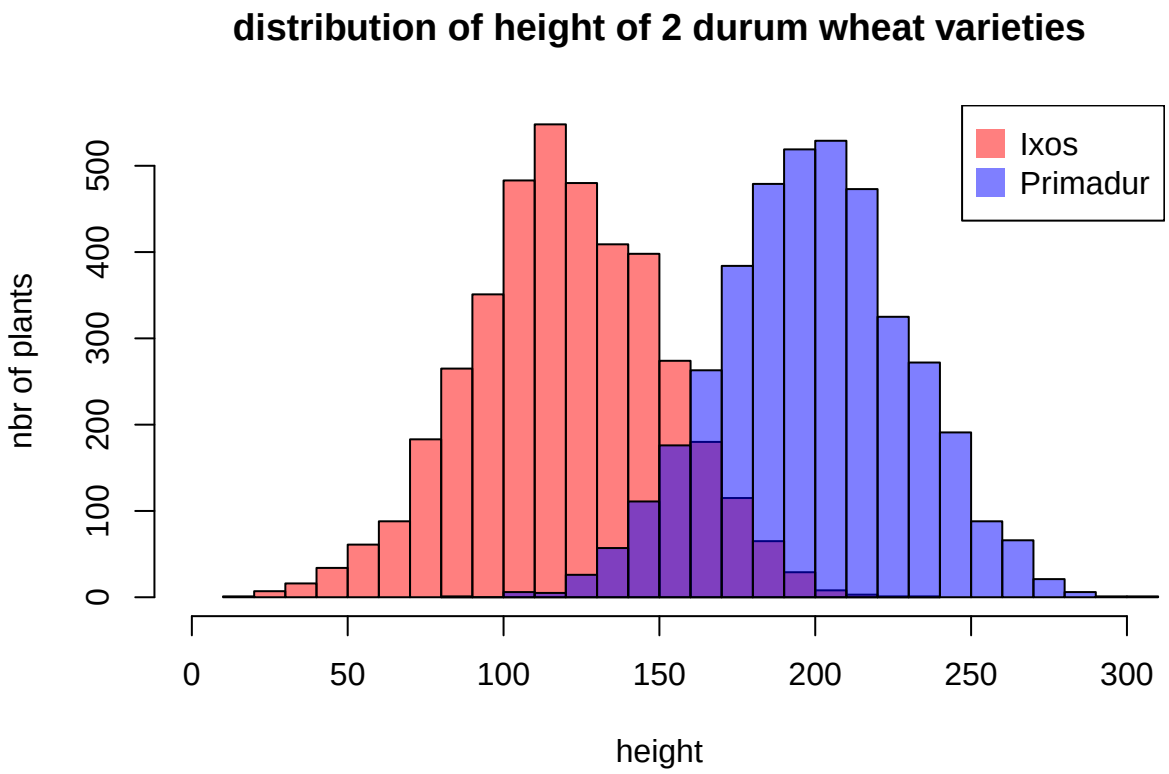


Diagrama de Caja

Es probablemente el tipo de gráfico más utilizado para comparar la distribución de varios grupos. Sin embargo, debes tener en cuenta que la distribución de datos se esconde detrás de cada cuadro. Por ejemplo, una distribución normal podría verse exactamente igual que una distribución bimodal.

```
# Load ggplot2
library(ggplot2)

# The mtcars dataset is natively available
# head(mtcars)

# A really basic boxplot.
ggplot(mtcars, aes(x=as.factor(cyl), y=mpg)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("cyl")
```

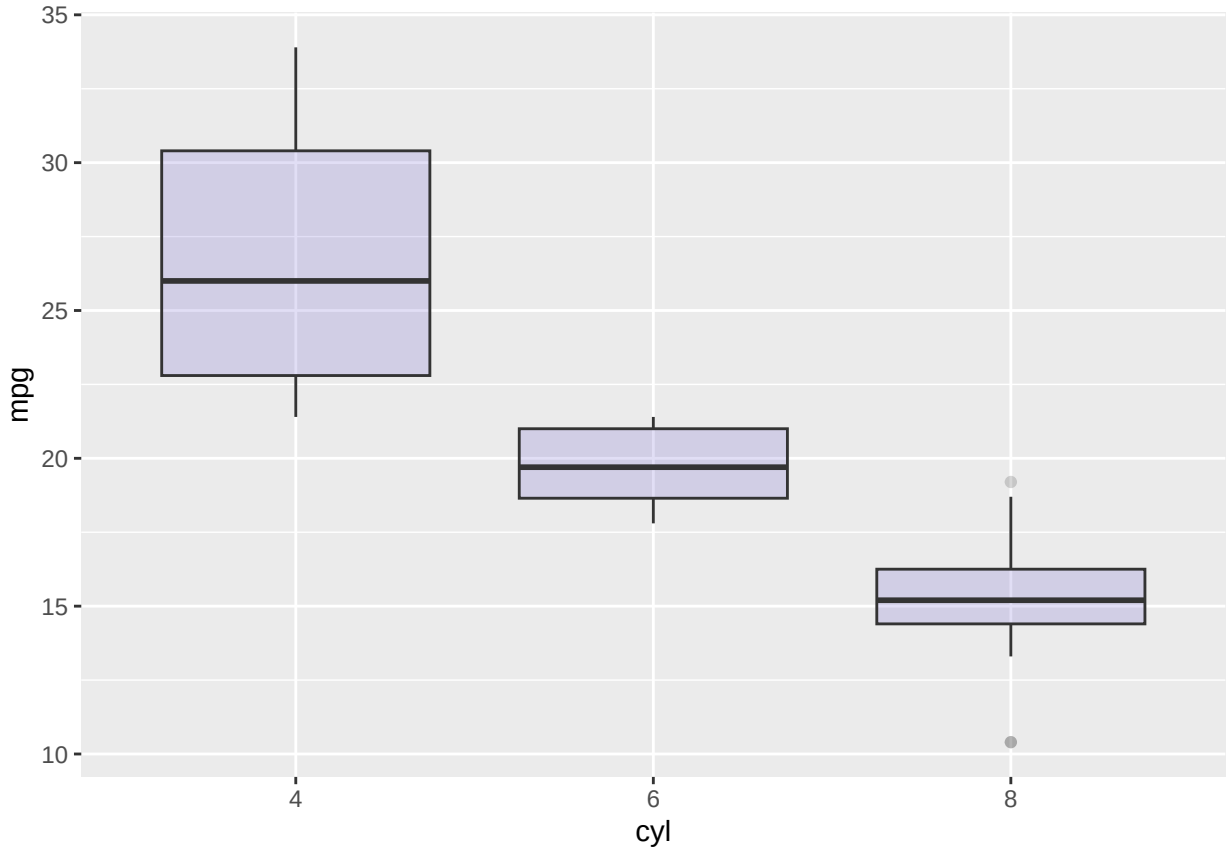


Gráfico de Dispersión

La capacidad de asignar una variable a características de marcador. Aquí, el marcador color depende de su valor en el campo llamado species en el marco de datos de entrada.

Tenga en cuenta que la leyenda se crea automáticamente.

```
# Create data
data = data.frame(
  x=seq(1:100) + 0.1*seq(1:100)*sample(c(1:10) , 100 , replace=T),
  y=seq(1:100) + 0.2*seq(1:100)*sample(c(1:10) , 100 , replace=T)
)

# Basic scatterplot
plot(x=data$x, y=data$y)
```

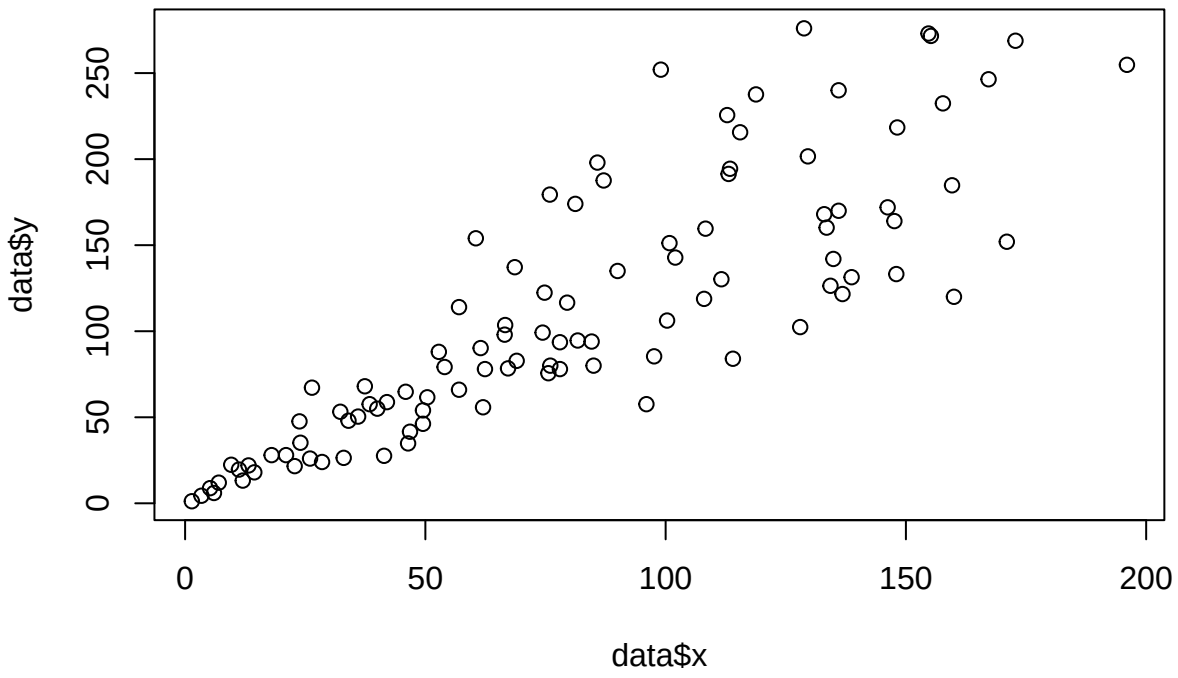
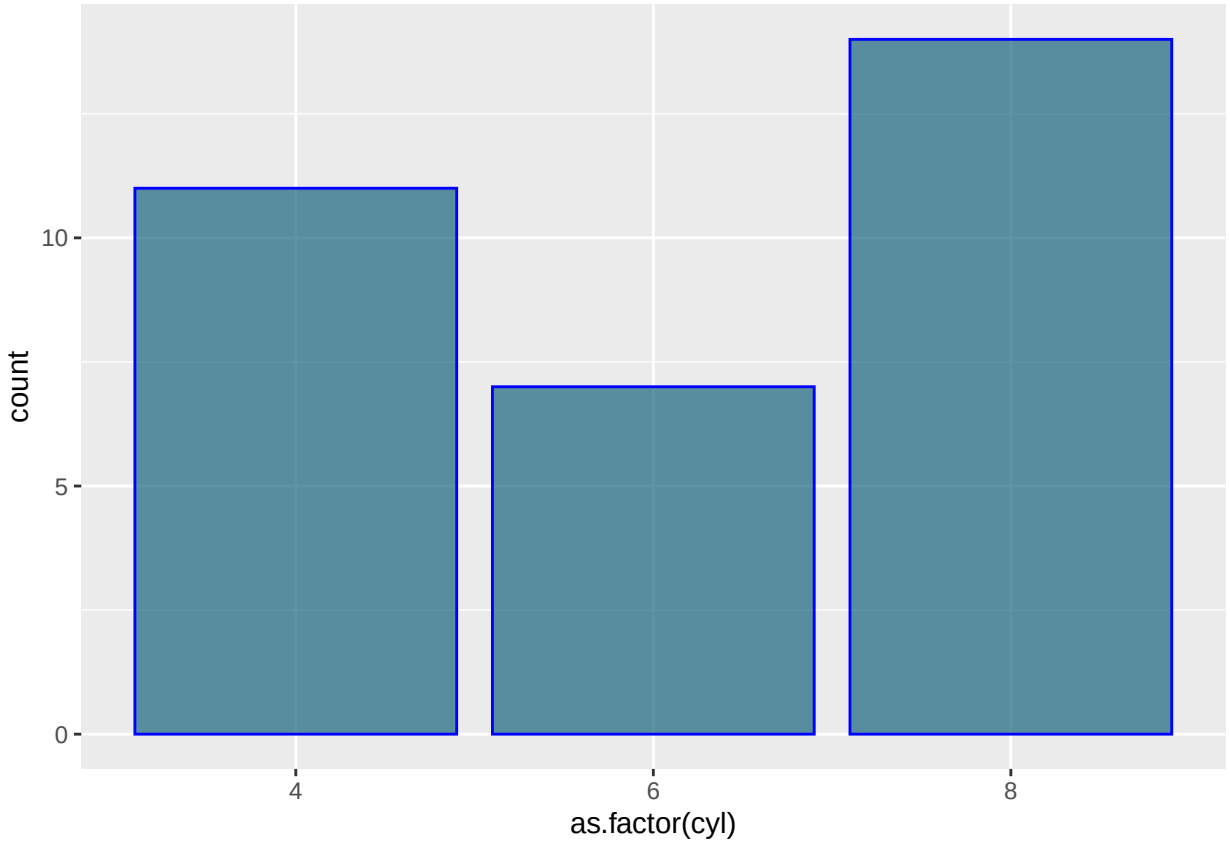


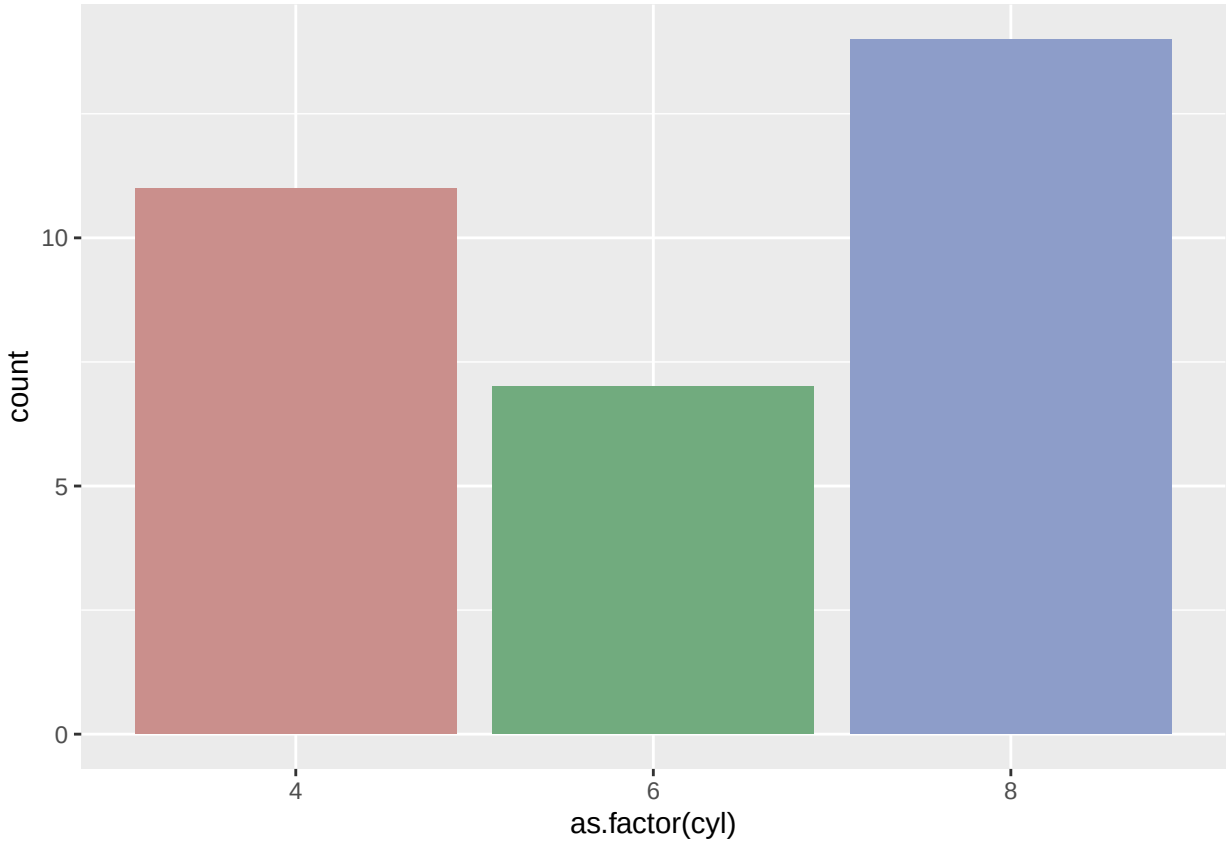
Diagrama de Barras

A continuación se muestran algunos métodos diferentes para controlar los colores de la barra. Tenga en cuenta que en este caso no es necesario utilizar una leyenda, ya que los nombres ya se muestran en el eje X.

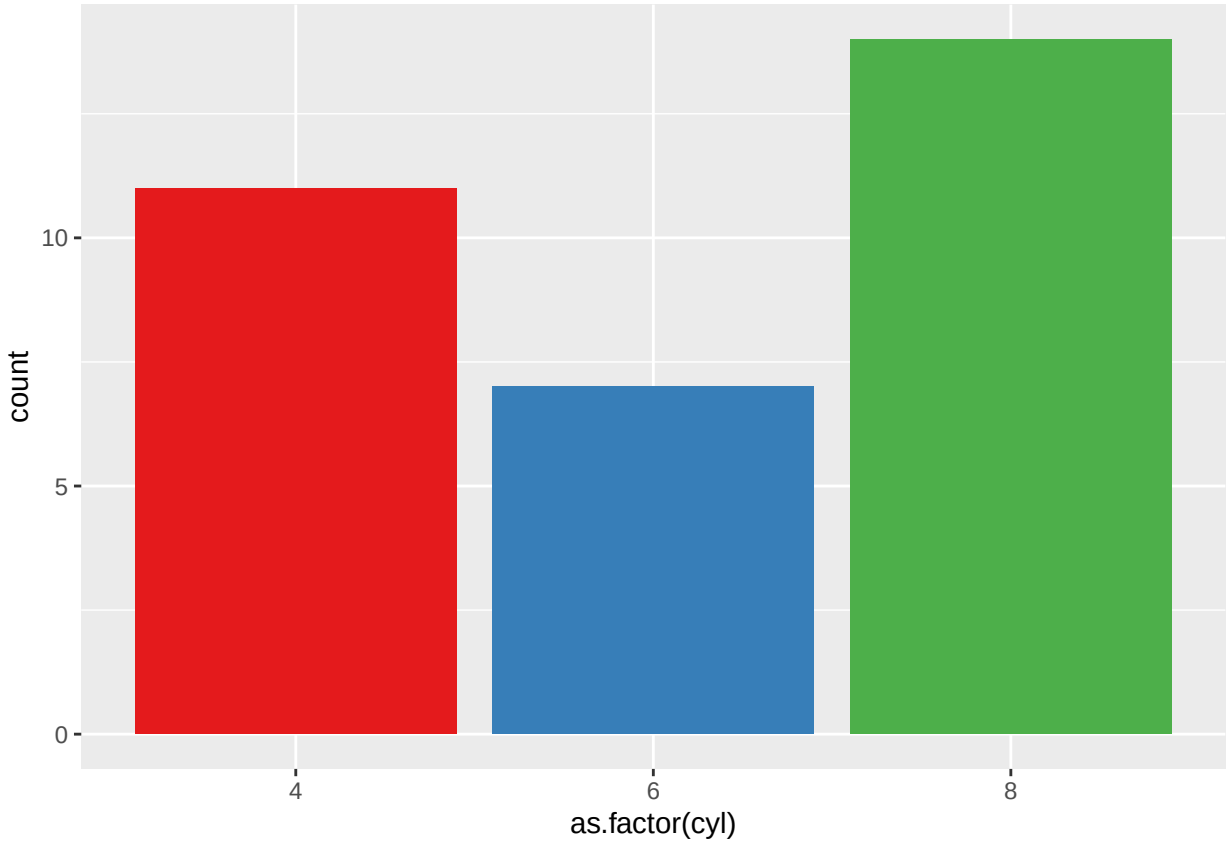
```
library(ggplot2)

# 1: uniform color. Color is for the border, fill is for the inside
ggplot(mtcars, aes(x=as.factor(cyl) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) )
```



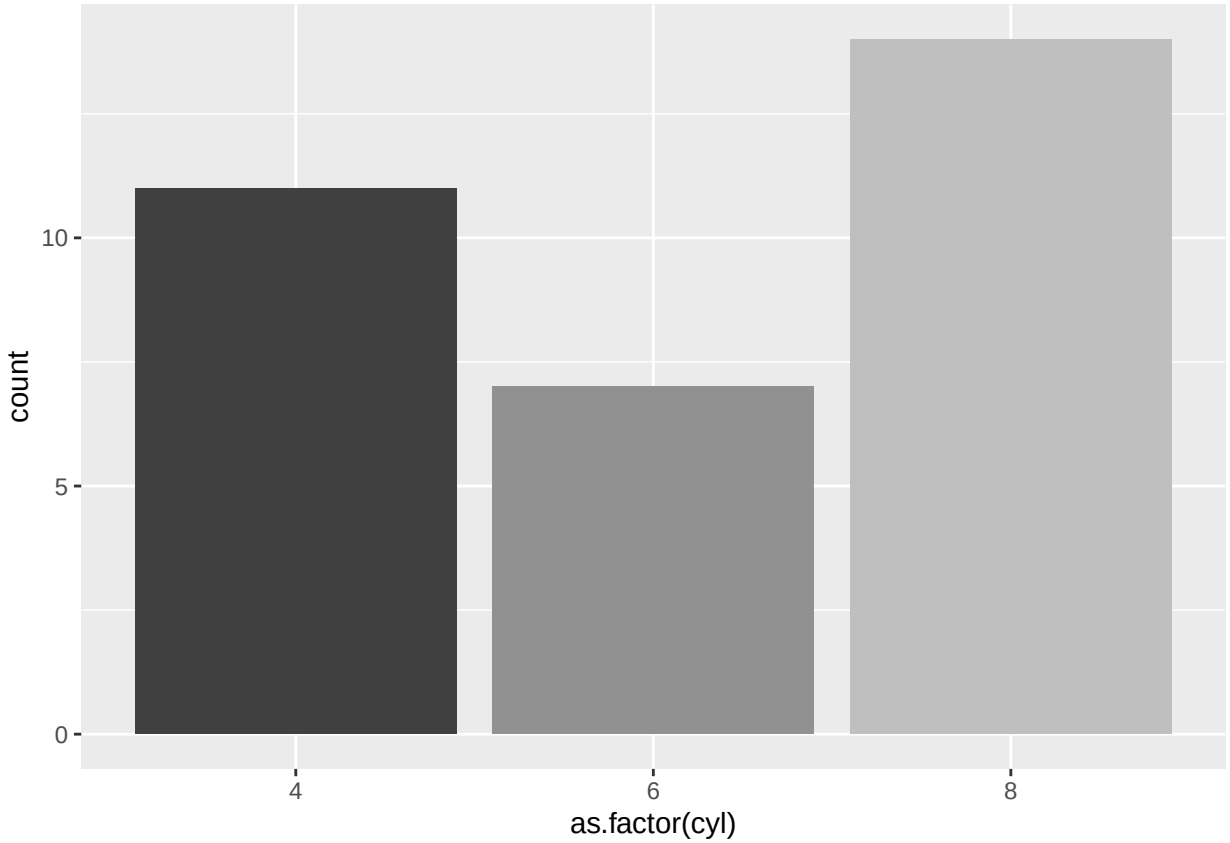

```
# 2: Using Hue
ggplot(mtcars, aes(x=as.factor(cyl), fill=as.factor(cyl) )) +
  geom_bar( ) +
  scale_fill_hue(c = 40) +
  theme(legend.position="none")
```




```
# 3: Using RColorBrewer
ggplot(mtcars, aes(x=as.factor(cyl), fill=as.factor(cyl) )) +
  geom_bar( ) +
  scale_fill_brewer(palette = "Set1") +
  theme(legend.position="none")
```




```
# 4: Using greyscale:
ggplot(mtcars, aes(x=as.factor(cyl), fill=as.factor(cyl) )) +
  geom_bar( ) +
  scale_fill_grey(start = 0.25, end = 0.75) +
  theme(legend.position="none")
```




```
# 5: Set manualy
ggplot(mtcars, aes(x=as.factor(cyl), fill=as.factor(cyl) )) +
  geom_bar( ) +
  scale_fill_manual(values = c("red", "green", "blue") ) +
  theme(legend.position="none")
```

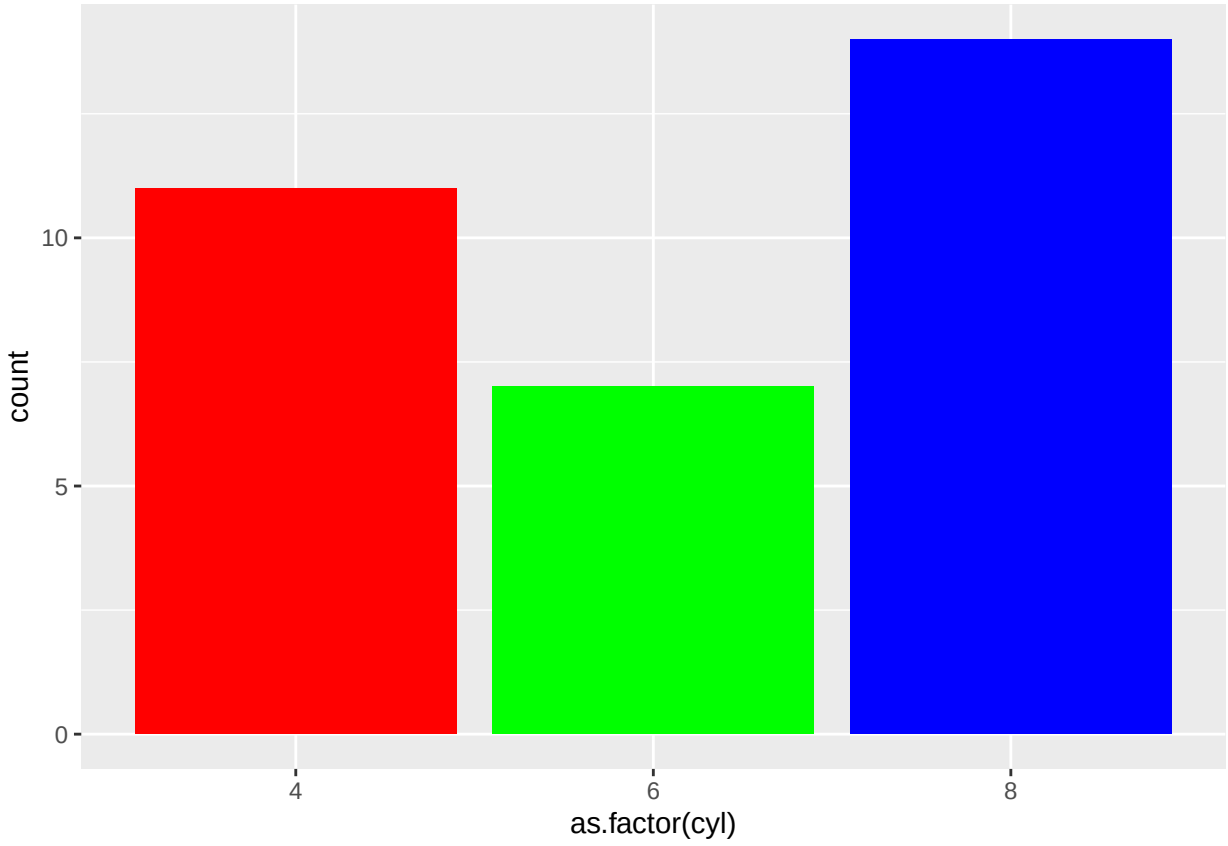


Gráfico de Linea

El marco de datos de entrada está compuesto por 3 columnas:

- Una variable numérica *ordenada* para el eje X.
- Otra variable numérica para el eje Y
- Una variable categórica que especifica el grupo de la observación.

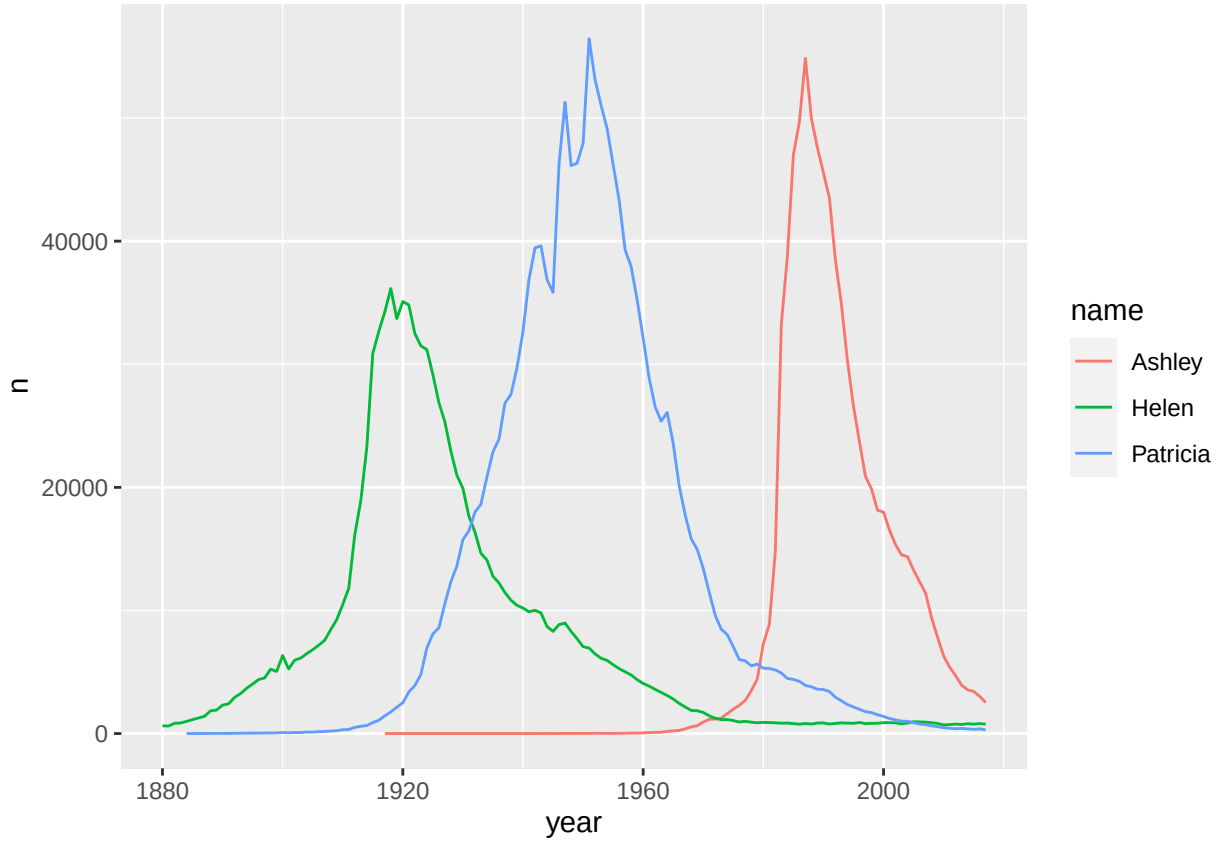
La idea es dibujar una línea por grupo. Esto se puede lograr especificando un color diferente para cada grupo.

```
# Libraries
library(ggplot2)
library(babynames) # provide the dataset: a dataframe called babynames
```

```
## Warning: package 'babynames' was built under R version 4.3.2
```

```
library(dplyr)

# Keep only 3 names
don <- babynames %>%
  filter(name %in% c("Ashley", "Patricia", "Helen")) %>%
  filter(sex=="F")

# Plot
don %>%
  ggplot( aes(x=year, y=n, group=name, color=name)) +
```

```
  geom_line()
```


40000
20000
0
n
1880
1920
1960
2000
year
name
Ashley
Helen
Patricia