

# Ejercicio 1.3: Graficos y Tablas

Silvia Pineda

## Índice

1. Haz un histograma de las variables cuantitativas y saca los estadísticos de las mismas. ¿Representarías algunas de estas variables con un diagrama de barras `barplot()`? Hazlo si la respuesta es Si. . . . . 1
2. Haz un gráfico para las variables cualitativas. ¿Están equilibradas las categorías de las variables cualitativas? Ayudate de los gráficos anteriores y haz unas tablas de frecuencias relativas para contestar. . . . . 10
3. Representa en un gráfico la variable salario por genero y edad por genero, luego haz los estadísticos `describeBy()` . . . . . 15

Para leer un csv con la modificación de los factores que ya hicimos hay que añadirle el argumento `stringsAsFactors`

```
datos<-read.csv("estres_corregido.csv",stringsAsFactors = TRUE)
```

**1. Haz un histograma de las variables cuantitativas y saca los estadísticos de las mismas. ¿Representarías algunas de estas variables con un diagrama de barras `barplot()`? Hazlo si la respuesta es Si.**

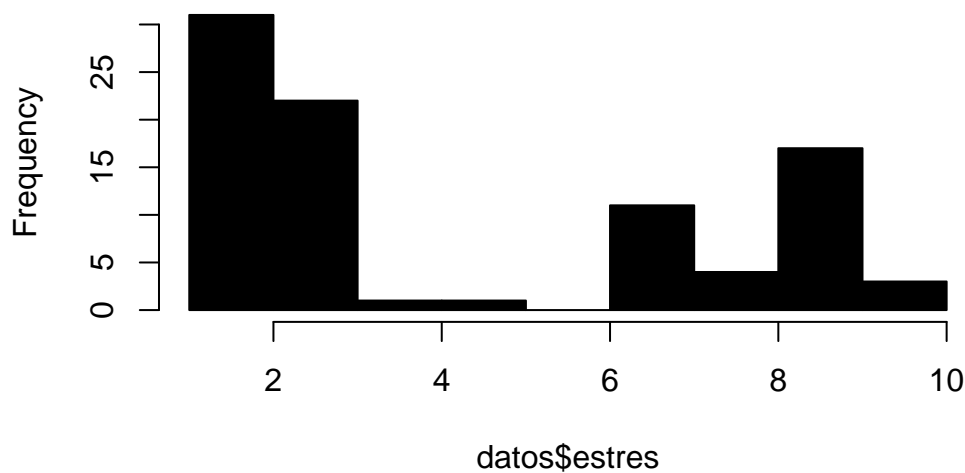
```
summary(datos)
```

X	ID	estres	grupo
Min. : 1.00	Min. : 1.00	Min. : 1.000	administrativo:18
1st Qu.:23.25	1st Qu.:23.25	1st Qu.: 2.000	enfermero :21
Median :45.50	Median :45.50	Median : 3.000	médico :23
Mean :45.50	Mean :45.50	Mean : 4.756	otros :28
3rd Qu.:67.75	3rd Qu.:67.75	3rd Qu.: 8.000	
Max. :90.00	Max. :90.00	Max. :10.000	

edad	peso	genero	sueldo	contrato
Min. :23.00	Min. :58.00	hombre:54	Min. :19000	fijo :38
1st Qu.:29.00	1st Qu.:67.00	mujer :36	1st Qu.:22000	temporal:52
Median :31.00	Median :72.00		Median :24000	
Mean :31.46	Mean :72.04		Mean :23633	
3rd Qu.:34.00	3rd Qu.:76.00		3rd Qu.:25000	
Max. :39.00	Max. :89.00		Max. :28000	
jornada	car_fisica	car_psiquica	ruido	
completa:44	Min. :1.000	Min. :2.000	Min. :0.000	
parcial :46	1st Qu.:2.000	1st Qu.:3.000	1st Qu.:2.000	
	Median :2.000	Median :3.000	Median :3.000	
	Mean :2.444	Mean :3.422	Mean :2.822	
	3rd Qu.:3.000	3rd Qu.:4.000	3rd Qu.:4.000	
	Max. :5.000	Max. :5.000	Max. :6.000	
car_fisica_corregida	expo_total	estres_grupo		
Min. :1.000	Min. : 3.000	alto :24		
1st Qu.:3.000	1st Qu.: 7.000	bajo :30		
Median :3.000	Median : 9.000	medio:35		
Mean :2.856	Mean : 8.689	NA's : 1		
3rd Qu.:3.000	3rd Qu.:11.000			
Max. :5.000	Max. :14.000			

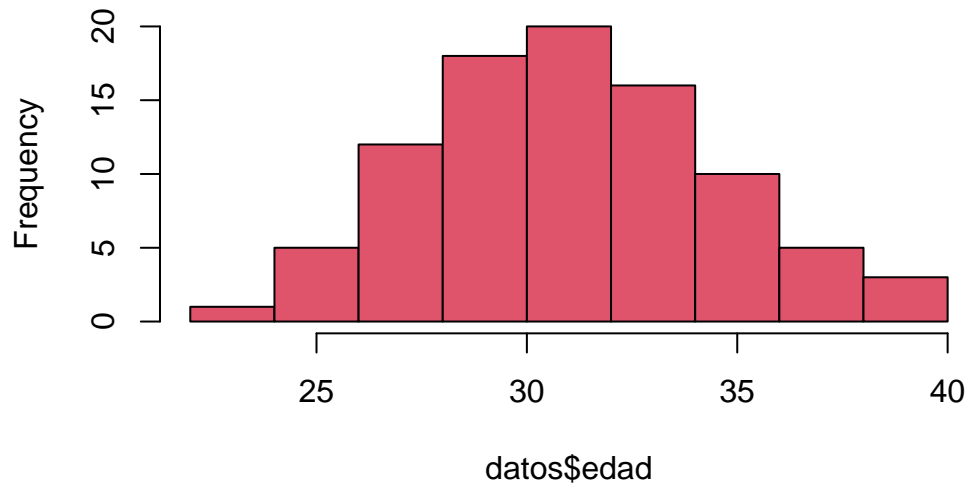
```
hist(datos$estres,col=1)
```

**Histogram of datos\$estres**



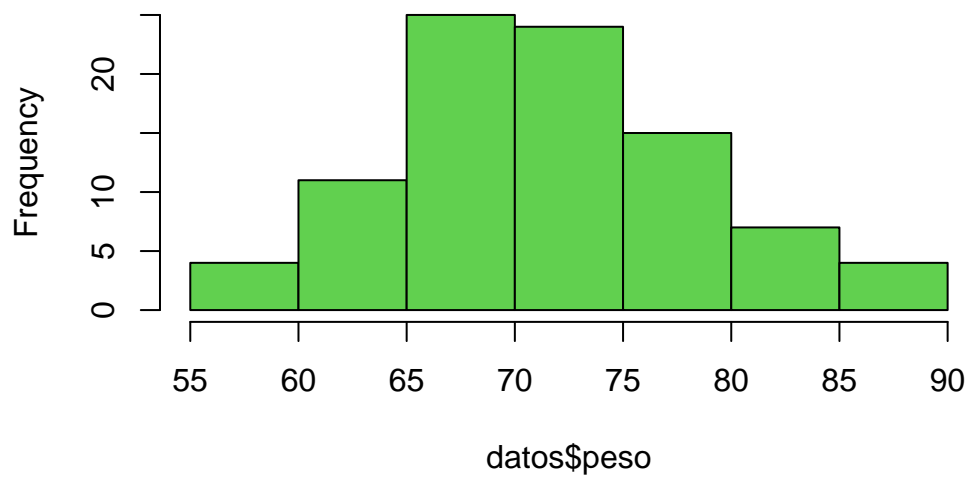
```
hist(datos$edad,col=2)
```

**Histogram of datos\$edad**

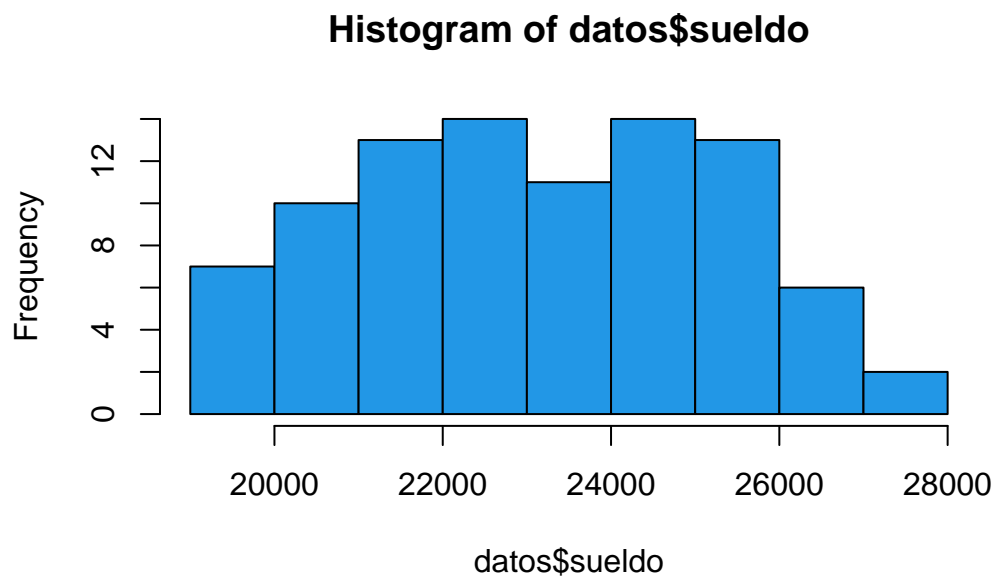


```
hist(datos$peso,col=3)
```

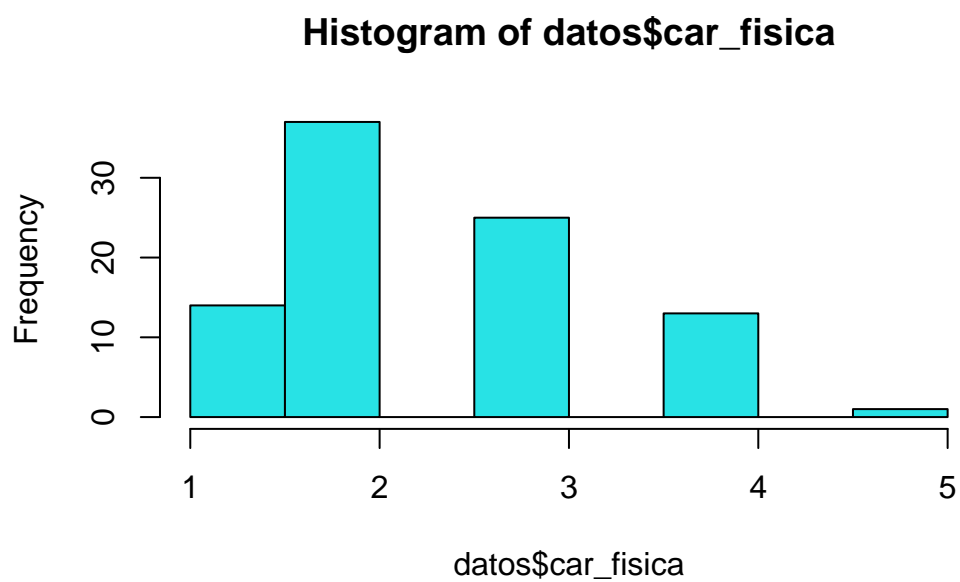
**Histogram of datos\$peso**



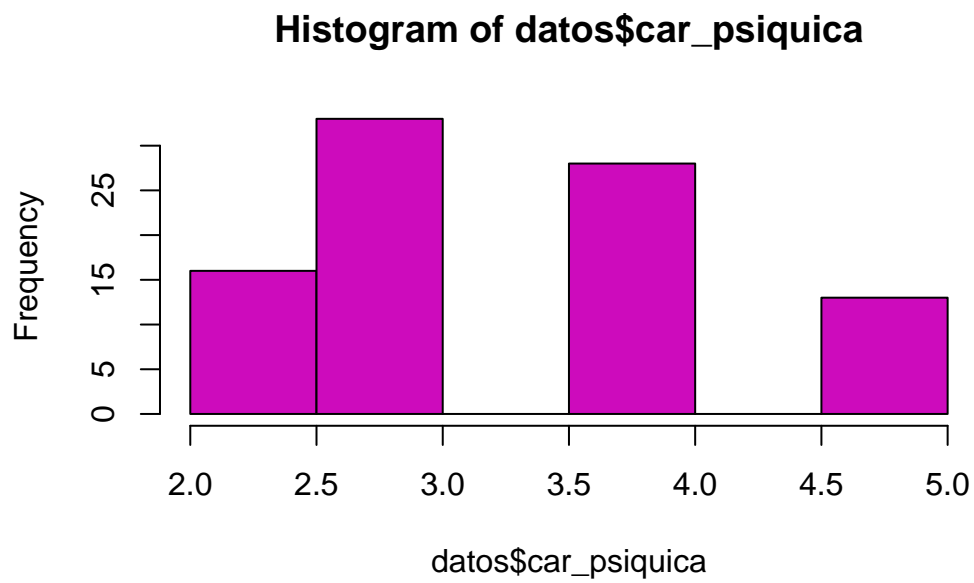
```
hist(datos$sueldo,col=4)
```



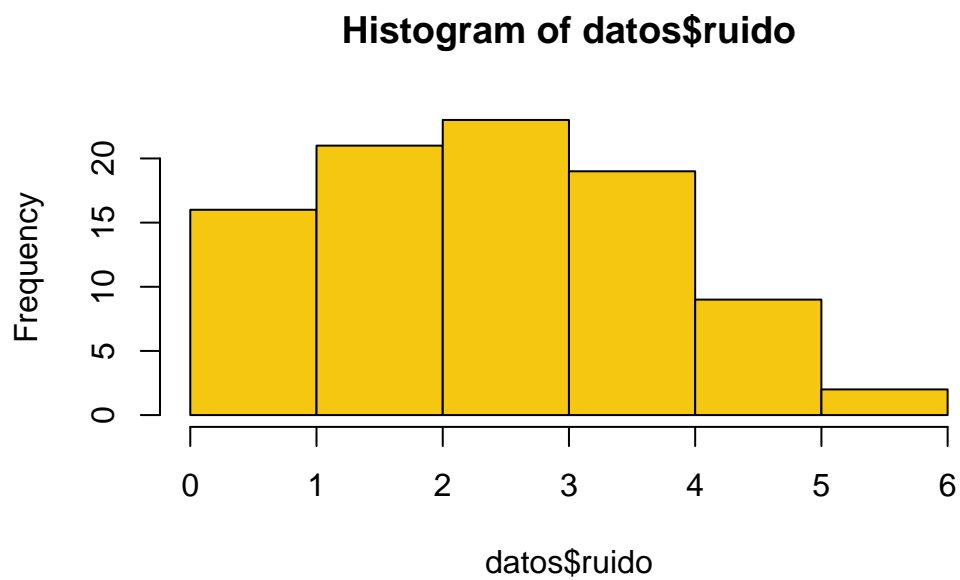
```
hist(datos$car_fisica,col=5)
```



```
hist(datos$car_psiquica,col=6)
```



```
hist(datos$ruido,col=7)
```



```
##Con ggplot
```

```
library(ggplot2)
```

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
v dplyr      1.2.0      v readr      2.1.6
```

```
v forcats    1.0.1      v stringr    1.6.0
```

```
v lubridate  1.9.4      v tibble     3.3.1
```

```
v purrr      1.2.1      v tidyr      1.3.2
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()     masks stats::lag()
```

```
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
datos |>
```

```
  select(where(is.numeric)) |>
```

```
  pivot_longer(everything(), names_to = "variable", values_to = "valor") |>
```

```
  ggplot(aes(x = valor, fill = variable)) +
```

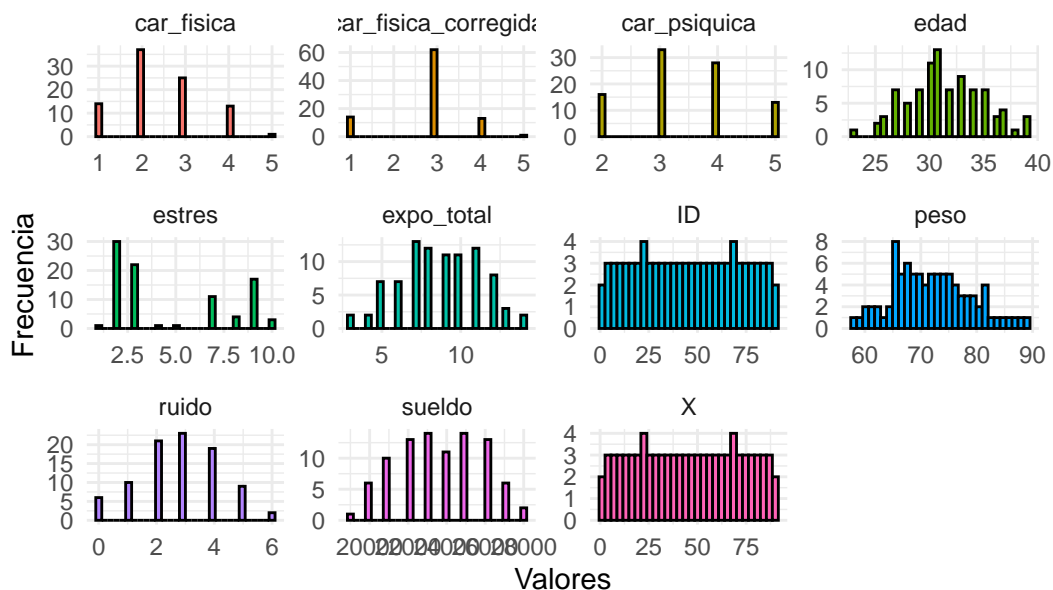
```
  geom_histogram(color = "black", bins = 30, show.legend = FALSE) +
```

```
  facet_wrap(~variable, scales = "free") +
```

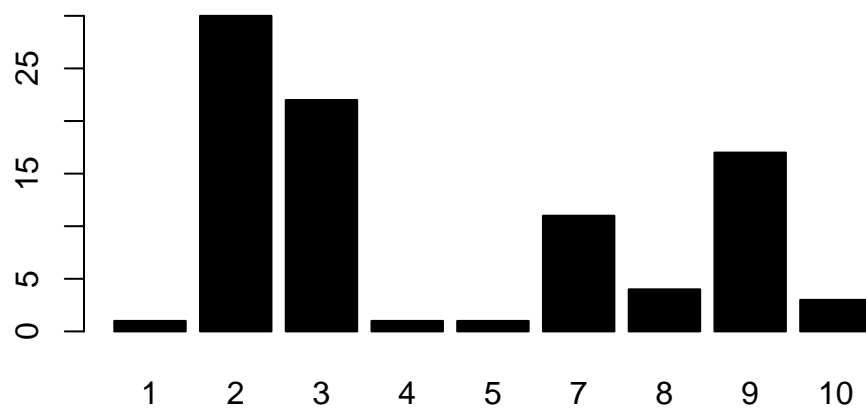
```
  labs(title = "Histogramas de variables numéricas", x = "Valores", y = "Frecuencia") +
```

```
  theme_minimal()
```

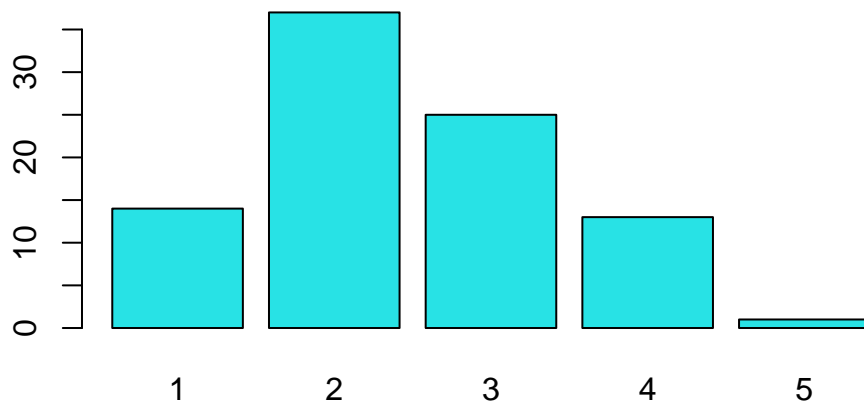
## Histogramas de variables numéricas



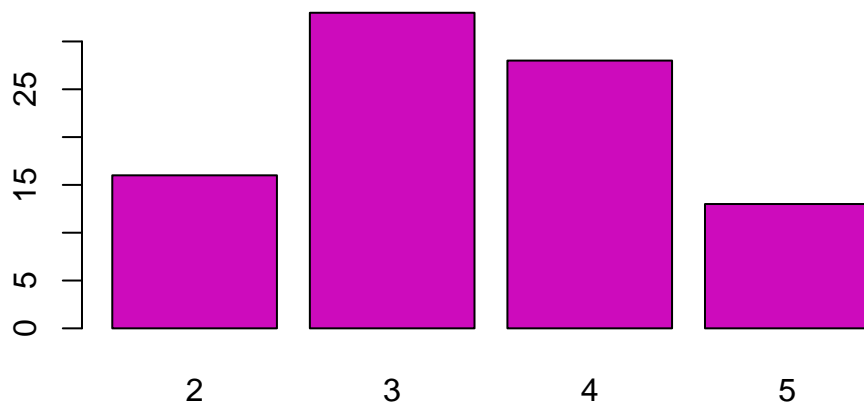
```
##Variables cuantitativas discretas  
barplot(table(datos$estres),col=1)
```



```
barplot(table(datos$car_fisica),col=5)
```

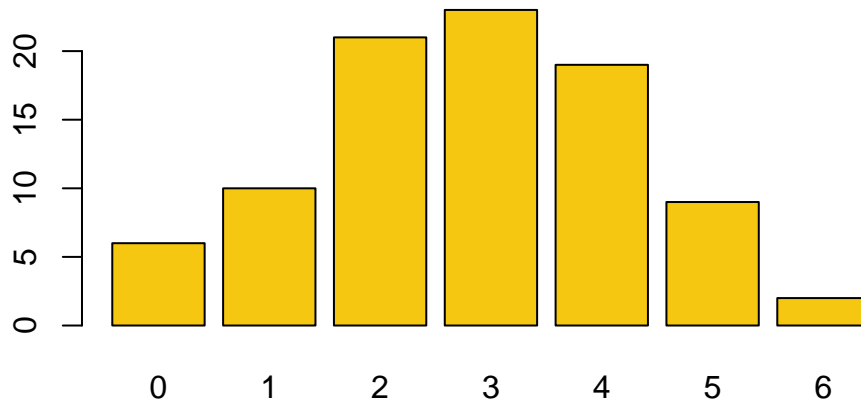


```
barplot(table(datos$car_psiquica),col=6)
```



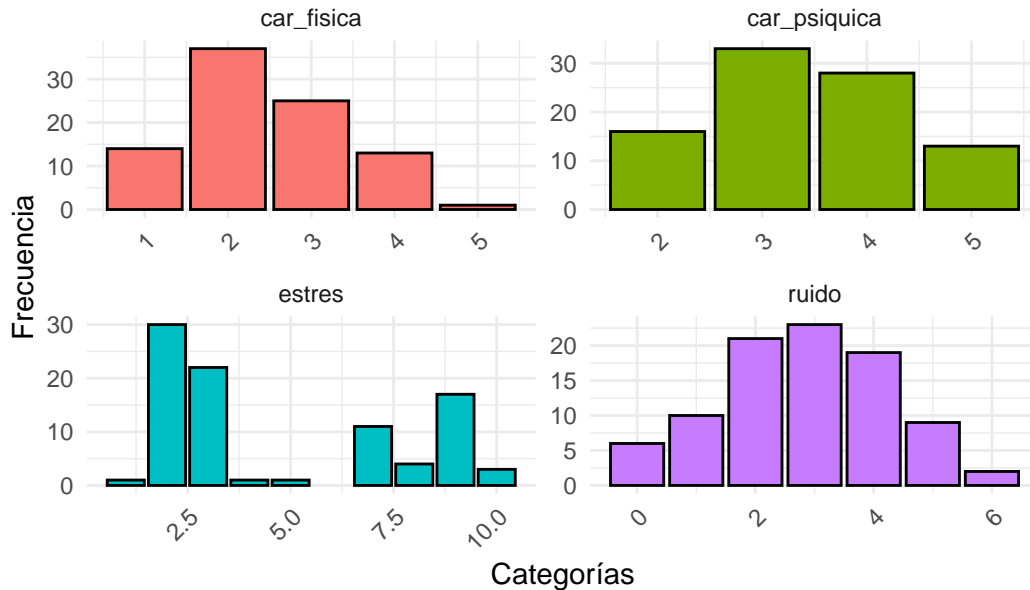


```
barplot(table(datos$ruido),col=7)
```



```
###Con ggplot
datos |>
  select(estres, car_fisica, car_psiquica, ruido) |>
  pivot_longer(everything(), names_to = "variable", values_to = "valor") |>
  ggplot(aes(x = valor, fill = variable)) +
  geom_bar(color = "black") +
  facet_wrap(~variable, scales = "free", ncol = 2) +
  labs(
    title = "Distribución de variables categóricas",
    x = "Categorías",
    y = "Frecuencia"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

## Distribución de variables categóricas



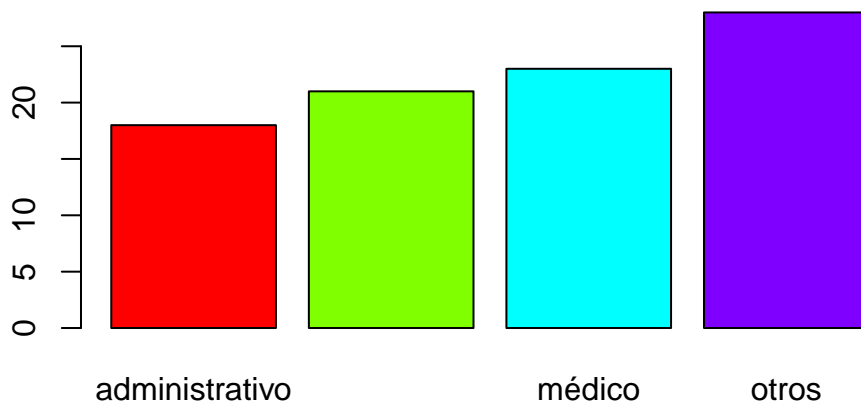
Las variables cuantitativas discretas que están medidas en una escala numérica suelen quedar mejor representadas con un diagrama de barras donde cada unidad numérica corresponde a una barra. En este caso son: `car_fisica`, `car_psiquica`, `estres` y `ruido`.

En el caso de estrés por ejemplo, el histograma no nos permite realmente ver como se distribuye la variable, ya que en el histograma parece que la mayoría tiene estrés de 1 cuando en realidad es de 2,3.

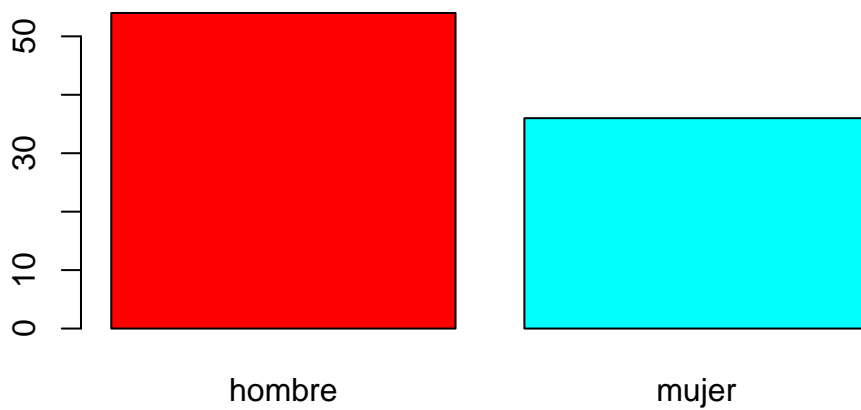
En el caso de ruido, el barplot nos permite como se distribuye más hacia una normal, algo que en el histograma no se apreciaba.

**2. Haz un gráfico para las variables cualitativas. ¿Están equilibradas las categorías de las variables cualitativas? Ayudate de los gráficos anteriores y haz unas tablas de frecuencias relativas para contestar.**

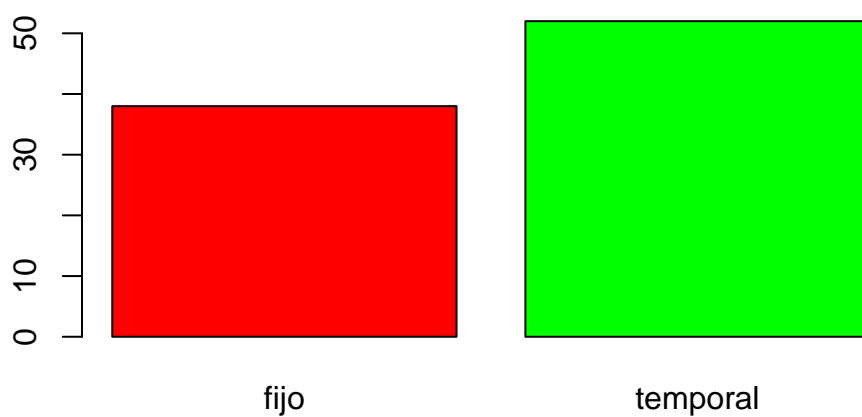
```
barplot(table(datos$grupo),col=rainbow(4))
```



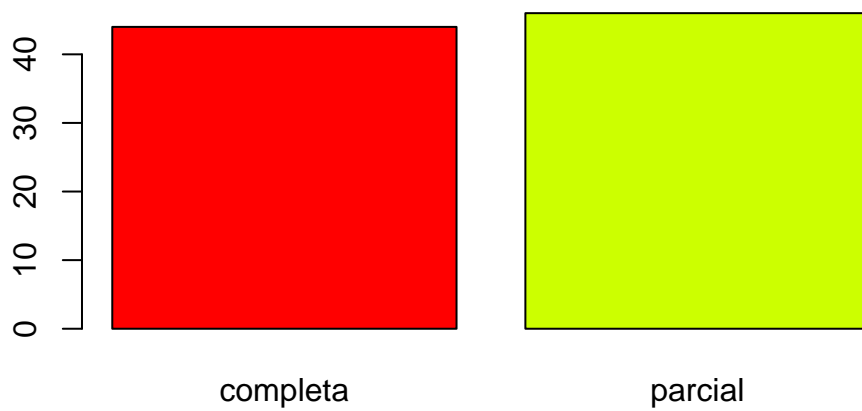
```
barplot(table(datos$genero),col=rainbow(2))
```



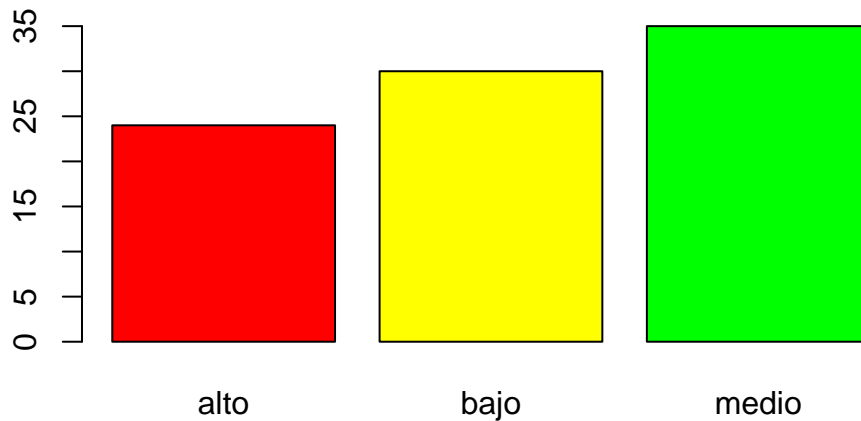
```
barplot(table(datos$contrato),col=rainbow(3))
```



```
barplot(table(datos$jornada),col=rainbow(5))
```

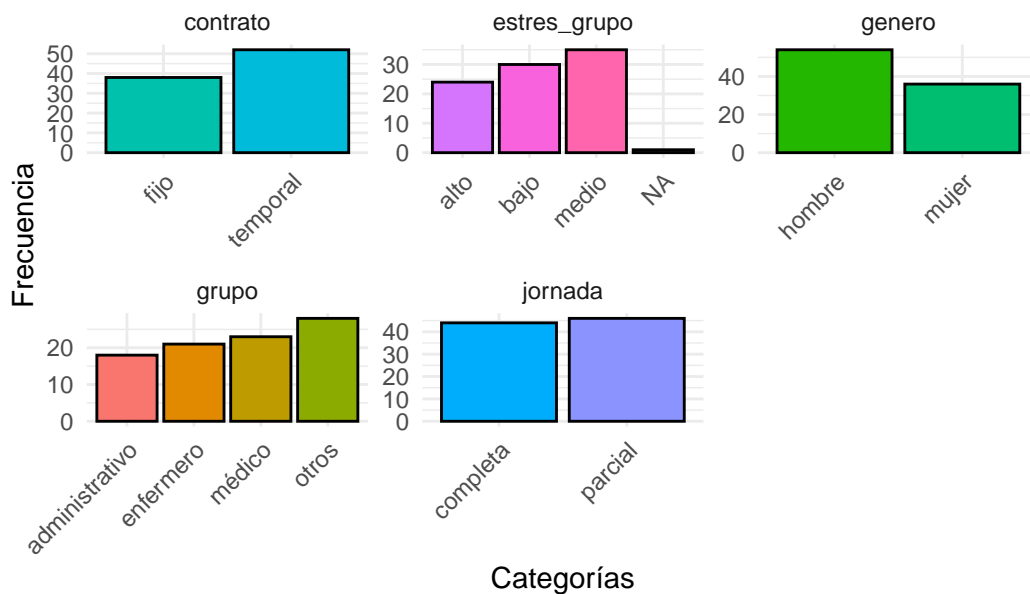


```
barplot(table(datos$estres_grupo), col=rainbow(6))
```



```
##Con ggplot
datos %>%
  select(grupo, genero, contrato, jornada, stres_grupo) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
  group_by(variable, valor) %>%
  summarise(frecuencia = n(), .groups = "drop") %>%
  ggplot(aes(x = valor, y = frecuencia, fill = valor)) +
  geom_col(color = "black") +
  facet_wrap(~variable, scales = "free", ncol = 3) +
  labs(
    title = "Distribución de Variables Categóricas",
    x = "Categorías",
    y = "Frecuencia"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

## Distribución de Variables Categóricas



```
prop.table(table(datos$grupo))
```

administrativo	enfermero	médico	otros
0.2000000	0.2333333	0.2555556	0.3111111

```
prop.table(table(datos$genero))
```

hombre	mujer
0.6	0.4

```
prop.table(table(datos$contrato))
```

fijo	temporal
0.4222222	0.5777778

```
prop.table(table(datos$jornada))
```

```
    completa    parcial  
0.4888889 0.5111111
```

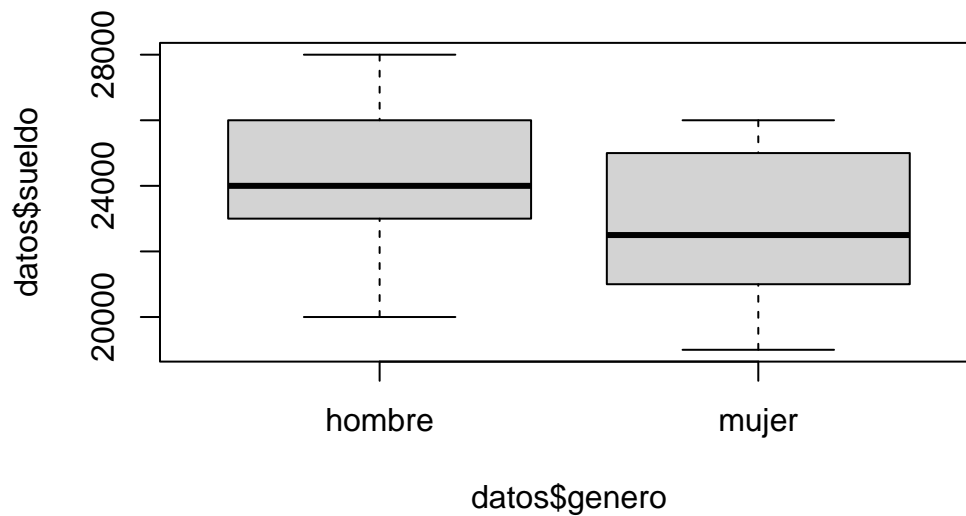
```
prop.table(table(datos$estres_grupo))
```

```
    alto    bajo    medio  
0.2696629 0.3370787 0.3932584
```

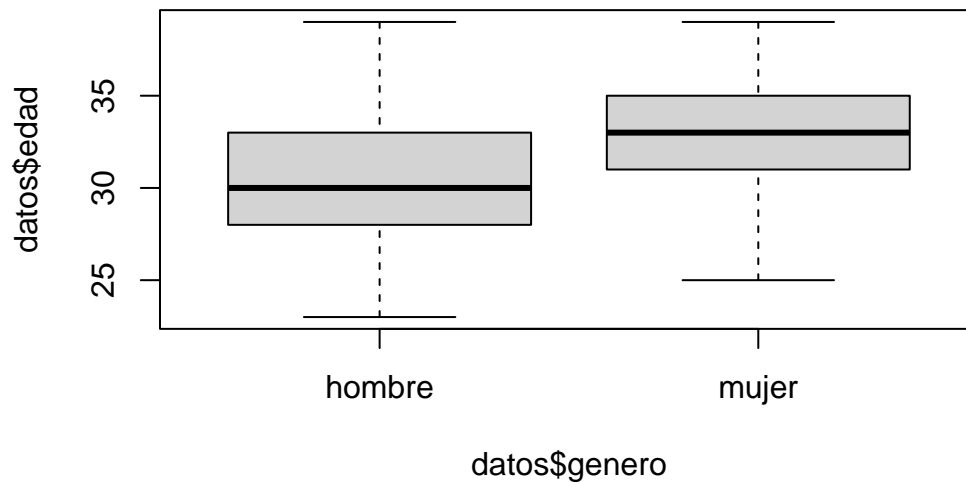
Todas las categorías de todas las variables están bastante bien representadas.

**3. Representa en un gráfico la variable salario por genero y edad por genero, luego haz los estadísticos describeBy()**

```
boxplot(datos$sueldo~datos$genero)
```

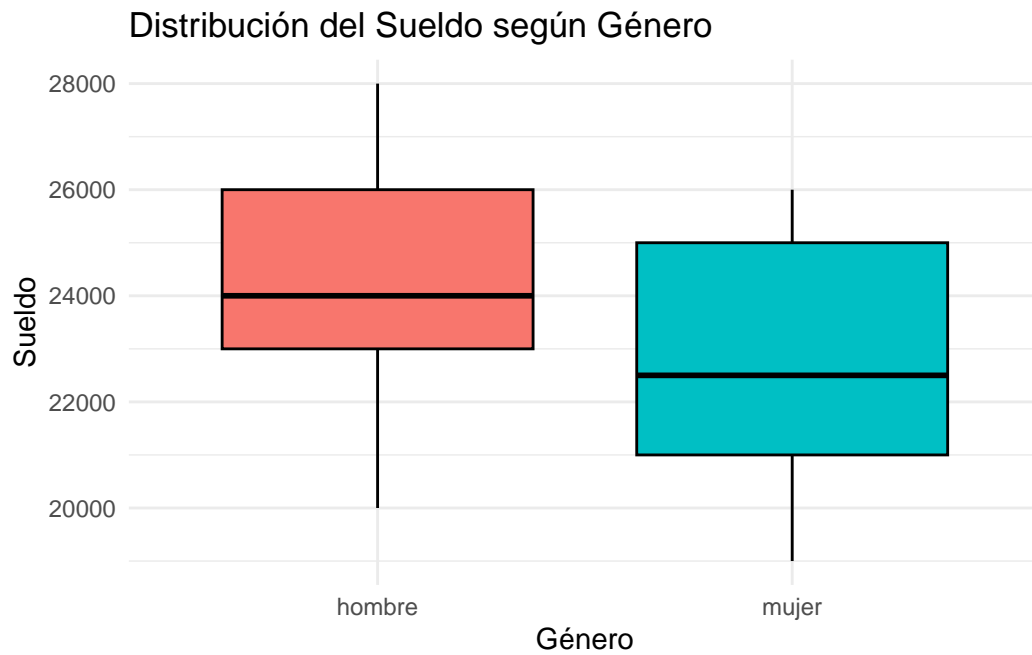


```
boxplot(datos$edad~datos$genero)
```

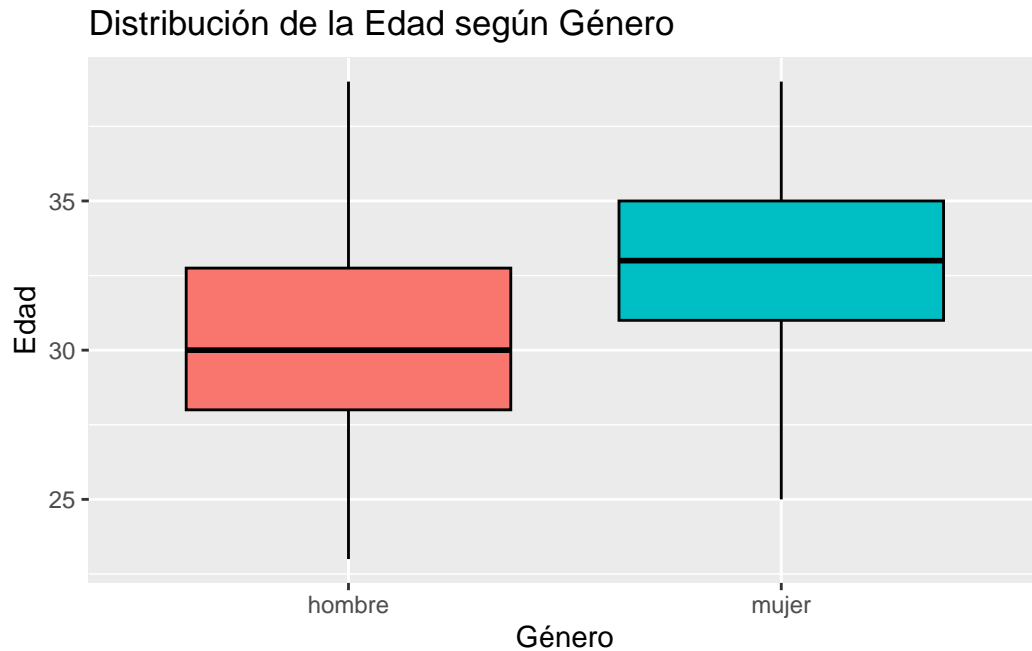


```
# Boxplot para salario según género
ggplot(datos, aes(x = genero, y = sueldo, fill = genero)) +
  geom_boxplot(color = "black") +
  labs(
    title = "Distribución del Sueldo según Género",
    x = "Género",
    y = "Sueldo"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```





```
# Boxplot para edad según género
ggplot(datos, aes(x = genero, y = edad, fill = genero)) +
  geom_boxplot(color = "black") +
  labs(
    title = "Distribución de la Edad según Género",
    x = "Género",
    y = "Edad"
  ) +
  theme(legend.position = "none")
```



```
library(psych)
```

Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

%+%, alpha

```
describeBy(datos$sueldo,datos$genero)
```

```
Descriptive statistics by group
group: hombre
  vars  n    mean      sd median trimmed   mad  min  max range  skew
X1    1 54 24259.26 2011.15  24000 24295.45 2965.2 20000 28000  8000 -0.08
      kurtosis    se
X1    -0.96 273.68
-----
group: mujer
  vars  n    mean      sd median trimmed   mad  min  max range  skew
```

```
X1      1 36 22694.44 2067.7 22500 22666.67 2223.9 19000 26000 7000 0.14
      kurtosis      se
X1      -1.25 344.62
```

```
describeBy(datos$edad,datos$genero)
```

Descriptive statistics by group

group: hombre

```
vars  n mean  sd median trimmed  mad min max range skew kurtosis  se
X1    1 54 30.5 3.4      30   30.34 2.97 23 39   16 0.39   -0.27 0.46
```

group: mujer

```
vars  n mean  sd median trimmed  mad min max range skew kurtosis  se
X1    1 36 32.89 3.08      33   32.9 2.97 25 39   14 -0.14    0.04 0.51
```

```
library(dplyr)
library(skimr)

datos_agrupados <- group_by(datos, genero)
skim(datos_agrupados, sueldo, edad)
```

Table 1: Data summary

Name	datos_agrupados
Number of rows	90
Number of columns	16
Column type frequency:	
numeric	2
Group variables	genero

### Variable type: numeric

skim_variable	genero	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
sueldo	hombre	0	1	24259.26	2011.15	20000	23000	24000	26000.00	28000	
sueldo	mujer	0	1	22694.44	2067.70	19000	21000	22500	25000.00	26000	
edad	hombre	0	1	30.50	3.40	23	28	30	32.75	39	

skim_variable	genero	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
edad	mujer	0	1	32.89	3.08	25	31	33	35.00	39	

Los hombres ganan más dinero, pero las mujeres son mayores