Practica1

Silvia Pineda

Instrucciones (leer antes de empezar)

Modifica dentro del documento Practica1.qmd tus datos personales (nombre y DNI) ubicados en la cabecera del archivo.

Asegúrate, ANTES de seguir editando el documento, que el archivo .qmd se renderiza correctamente y se genera el .html correspondiente en tu carpeta local de tu ordenador.

Los chunks (cajas de código) creados están o vacíos o incompletos, de ahí que la mayoría tengan la opción **#| eval: false**. Una vez que edites lo que consideres, debes ir cambiando cada chunck a **#| eval: true** (o quitarlo directamente) para que se ejecuten.

ENUNCIADO DE LA PRÁCTICA Para esta práctica vais a usar los datos del fichero Alumnos.csv que contienen información sobre los alumnos de un máster universitario. Este fichero consiste de 100 observaciones y 12 variables. La definición de las variables la tienes en el diccionario de variables del gitbook de la PRACTICA 1.

CARGA DE DATOS

```
#Asegúrate de estar en el directorio dónde has guardado los datos
datos<-read.csv("Alumnos.csv")
```

EJERCICIO 1: ¿Los nombres de las variables son consistentes? Modifícalos si no es así

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- 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 becor
```

```r
datos <-janitor::clean_names(datos)
colnames(datos)[2]<-"edad"
```

EJERCICIO 2: ¿El tipo de variable está declarado de forma correcta? Modifícalo si no es así

```r
str(datos)
```

```
'data.frame':   100 obs. of  12 variables:
 $ id              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ edad            : int  48 32 31 20 59 60 54 31 42 43 ...
 $ ingresos        : num  19100 14018 10382 17011 13582 ...
 $ genero          : chr  "f" "M" "m" "F" ...
 $ vivienda        : chr  "Amigos" "Amigos" "Amigos" "Familia" ...
 $ puntuacion      : int  8 6 9 7 9 6 7 8 6 6 ...
 $ altura          : num  169 152 154 175 152 ...
 $ peso            : num  76.5 64.6 65.6 63.2 70.9 ...
 $ tiempo_libre    : num  3.84 2.71 3.5 1.84 2.87 ...
 $ consumo_cafe    : chr  "Alto" "Moderado" "Alto" "Alto" ...
 $ horas_deporte   : num  2.54 2.616 2.617 0.308 2.369 ...
 $ horas_de_trabajo: num  9.82 8.28 5.22 6.27 7.67 ...
```

```r
columns_to_factor<-c("genero","vivienda","consumo_cafe")
datos[,columns_to_factor]<-
    lapply(datos[,columns_to_factor],as.factor)
str(datos)
```

```
'data.frame':   100 obs. of  12 variables:
```

```
 $ id              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ edad            : int  48 32 31 20 59 60 54 31 42 43 ...
 $ ingresos        : num  19100 14018 10382 17011 13582 ...
 $ genero          : Factor w/ 5 levels "f","F","m","M",..: 1 4 3 2 1 4 5 4 3 1 ...
 $ vivienda        : Factor w/ 4 levels "Amigos","familia",..: 1 1 1 3 1 3 1 1 1 1 ...
 $ puntuacion      : int  8 6 9 7 9 6 7 8 6 6 ...
 $ altura          : num  169 152 154 175 152 ...
 $ peso            : num  76.5 64.6 65.6 63.2 70.9 ...
 $ tiempo_libre    : num  3.84 2.71 3.5 1.84 2.87 ...
 $ consumo_cafe    : Factor w/ 4 levels "?","Alto","Bajo",..: 2 4 2 2 2 3 3 2 3 3 ...
 $ horas_deporte   : num  2.54 2.616 2.617 0.308 2.369 ...
 $ horas_de_trabajo: num  9.82 8.28 5.22 6.27 7.67 ...
```

EJERCICIO 3: ¿Todas las variables toman valores correctos? Si hay valores incorrectos haz que aparezcan como missing (NA) o si los puedes identificar, corrígelos.

```
table(datos$genero)
```

```
 f  F  m  M NB
 5 32  6 54  3
```

```
datos$genero<-car::recode(datos$genero,"'f'='F';'m'='M'")
table(datos$genero)
```

```
 F  M NB
37 60  3
```

```
table(datos$vivienda)
```

```
 Amigos familia Familia    Solo
     70       1      26       3
```

```
datos$vivienda<-car::recode(datos$vivienda,"'familia'='Familia'")
table(datos$vivienda)
```

```
 Amigos Familia    Solo
     70      27       3
```

```
table(datos$consumo_cafe)
```

```
       ?     Alto     Bajo Moderado
       3       33       35       29
```

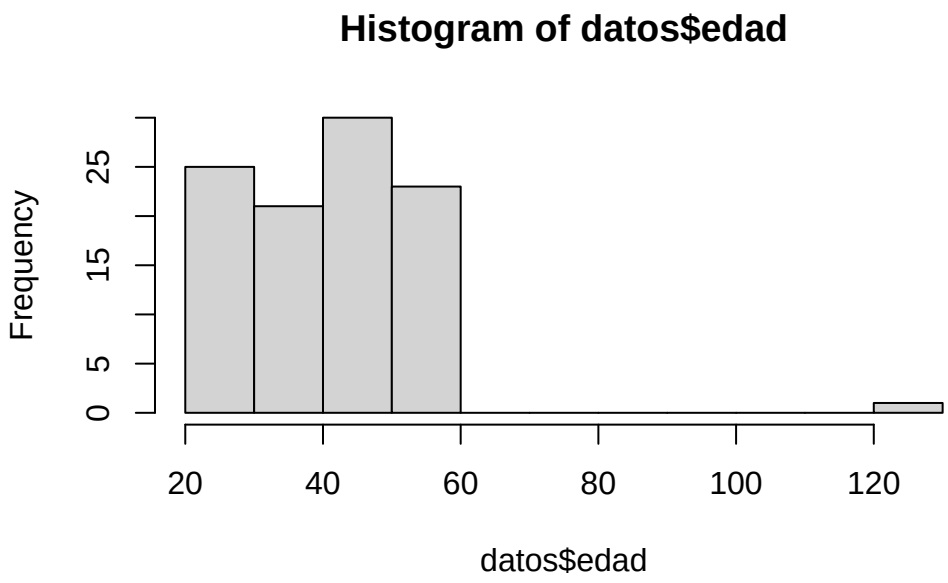
```
datos$consumo_cafe<-car::recode(datos$consumo_cafe,"'?'= NA")
table(datos$consumo_cafe)
```

```
    Alto     Bajo Moderado
      33       35       29
```

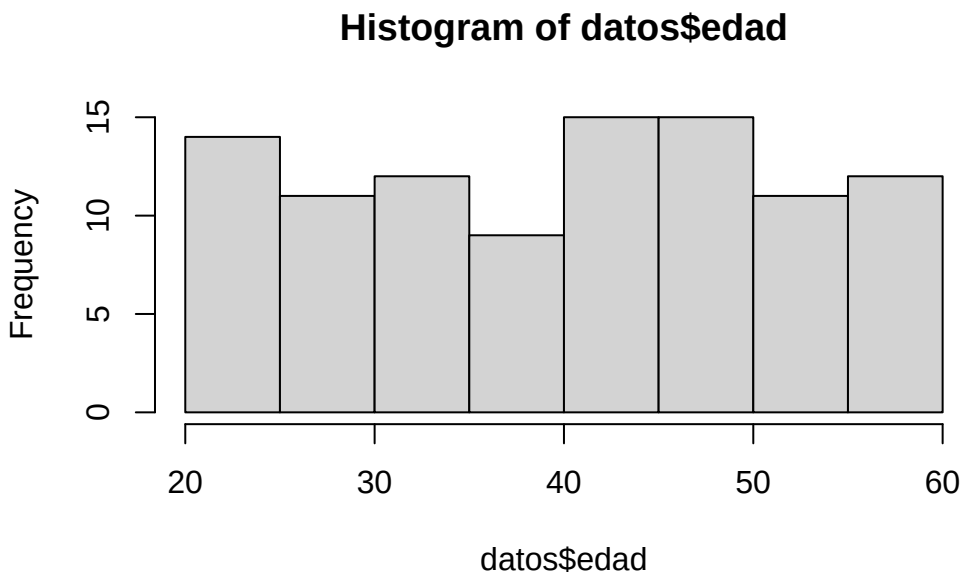
```
summary(datos)
```

```
       id              edad           ingresos     genero     vivienda
 Min.   :  1.00   Min.   : 20.00   Min.   :10382   F :37   Amigos :70
 1st Qu.: 25.75   1st Qu.: 30.75   1st Qu.:13678   M :60   Familia:27
 Median : 50.50   Median : 42.00   Median :14742   NB: 3   Solo   : 3
 Mean   : 50.50   Mean   : 41.33   Mean   :14941
 3rd Qu.: 75.25   3rd Qu.: 49.25   3rd Qu.:16230
 Max.   :100.00   Max.   :123.00   Max.   :21482
  puntuacion        altura           peso        tiempo_libre
 Min.   : 4.00   Min.   :143.5   Min.   :55.95   Min.   : 0.3983
 1st Qu.: 6.00   1st Qu.:156.5   1st Qu.:66.46   1st Qu.: 2.1621
 Median : 7.00   Median :162.5   Median :69.67   Median : 2.8961
 Mean   : 6.97   Mean   :164.0   Mean   :69.71   Mean   : 3.1026
 3rd Qu.: 8.00   3rd Qu.:168.8   3rd Qu.:72.90   3rd Qu.: 3.6140
 Max.   :10.00   Max.   :210.0   Max.   :82.15   Max.   :28.0000
  consumo_cafe horas_deporte      horas_de_trabajo
 Alto    :33   Min.   : -0.5079   Min.   : 4.150
 Bajo    :35   1st Qu.:  1.4059   1st Qu.: 6.663
 Moderado:29   Median :  2.0602   Median : 8.286
 NA's    : 3   Mean   : 31.9214   Mean   : 8.214
               3rd Qu.:  2.7917   3rd Qu.: 9.397
               Max.   :999.0000   Max.   :13.151
```

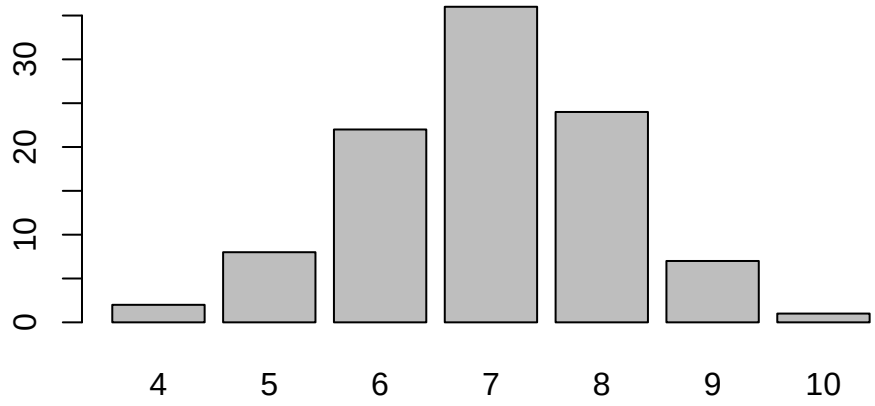
```r
hist(datos$edad)
```



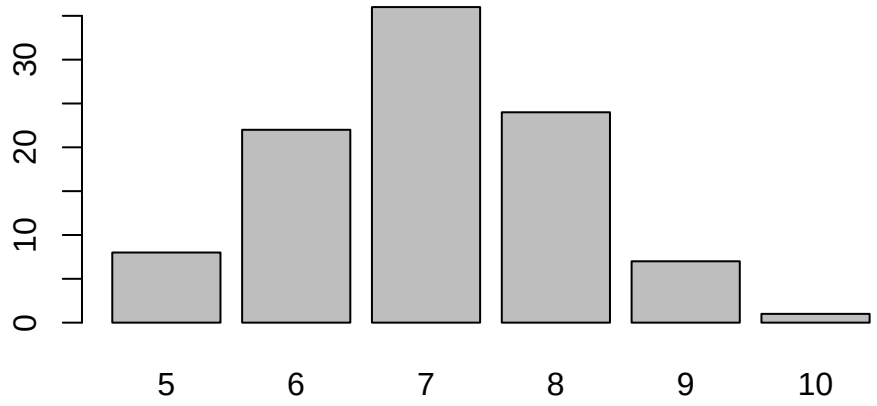

```r
datos$edad<-replace(datos$edad,datos$edad==123,NA)
hist(datos$edad)
```



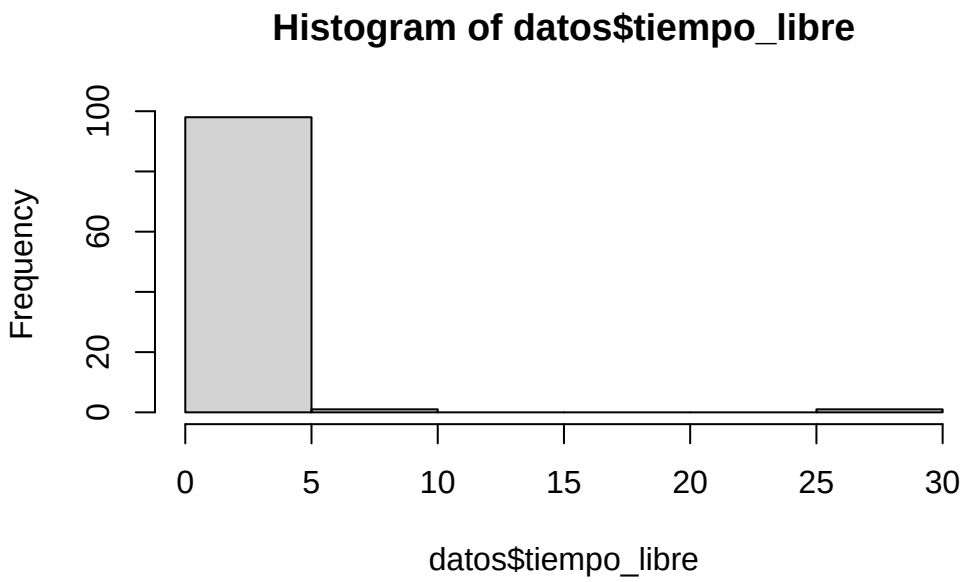

```r
barplot(table(datos$puntuacion))
```



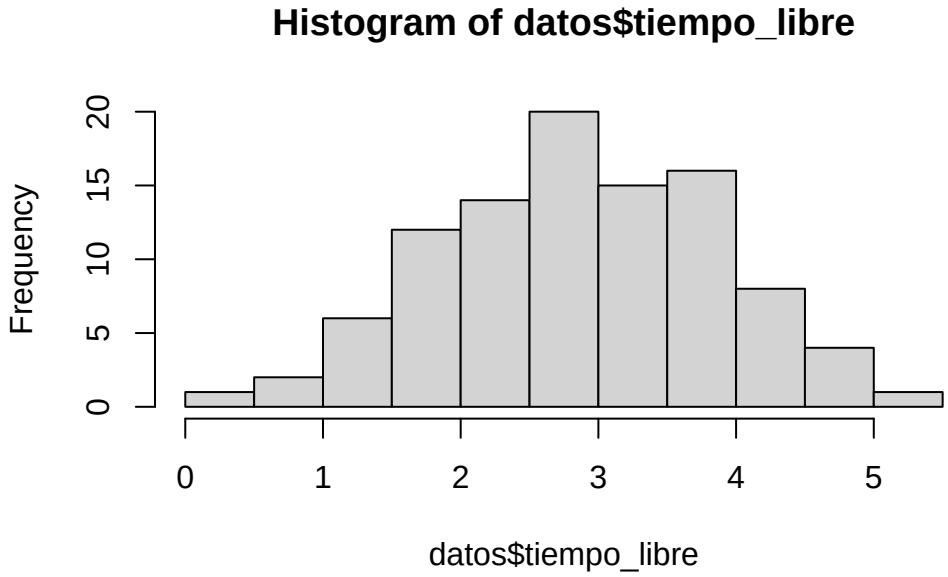
```r
datos$puntuacion<-replace(datos$puntuacion,datos$puntuacion<5,NA)
barplot(table(datos$puntuacion))
```



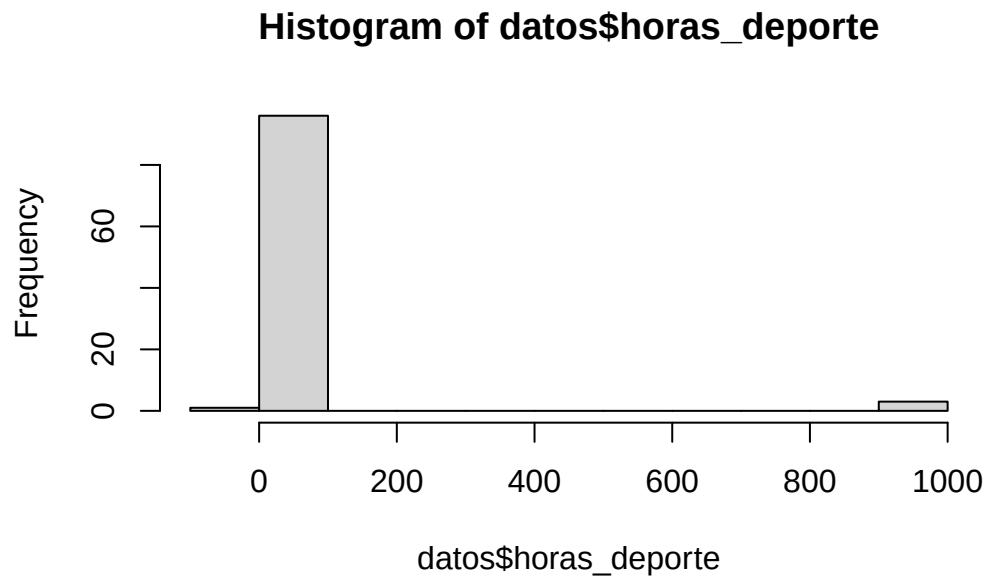

```r
hist(datos$tiempo_libre)
```



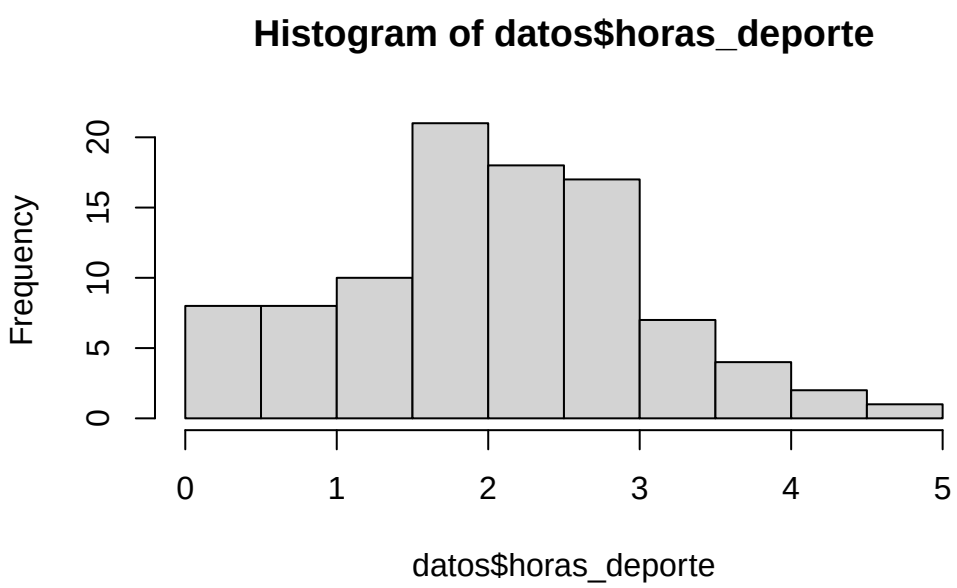

```
datos$tiempo_libre<-replace(datos$tiempo_libre,datos$tiempo_libre>24,NA)
hist(datos$tiempo_libre)
```



```r
hist(datos$horas_deporte)
```




```r
datos$horas_deporte<-replace(datos$horas_deporte,datos$horas_deporte==999,NA)
datos$horas_deporte<-replace(datos$horas_deporte,datos$horas_deporte<0,NA)
hist(datos$horas_deporte)
```



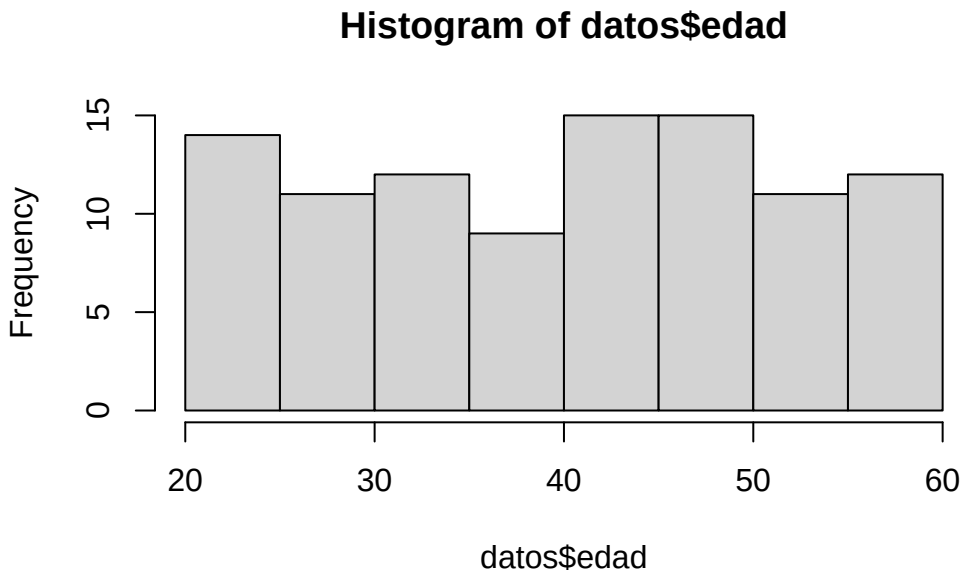

```
summary(datos)
```

```
       id              edad          ingresos     genero     vivienda
 Min.   :  1.00   Min.   :20.00   Min.   :10382   F :37   Amigos :70
 1st Qu.: 25.75   1st Qu.:30.50   1st Qu.:13678   M :60   Familia:27
 Median : 50.50   Median :42.00   Median :14742   NB: 3   Solo   : 3
 Mean   : 50.50   Mean   :40.51   Mean   :14941
 3rd Qu.: 75.25   3rd Qu.:49.00   3rd Qu.:16230
 Max.   :100.00   Max.   :60.00   Max.   :21482
                  NA's   :1
   puntuacion         altura           peso         tiempo_libre
 Min.   : 5.000   Min.   :143.5   Min.   :55.95   Min.   :0.3983
 1st Qu.: 6.000   1st Qu.:156.5   1st Qu.:66.46   1st Qu.:2.1584
 Median : 7.000   Median :162.5   Median :69.67   Median :2.8837
 Mean   : 7.031   Mean   :164.0   Mean   :69.71   Mean   :2.8511
 3rd Qu.: 8.000   3rd Qu.:168.8   3rd Qu.:72.90   3rd Qu.:3.5885
 Max.   :10.000   Max.   :210.0   Max.   :82.15   Max.   :5.0867
 NA's   :2                                        NA's   :1
   consumo_cafe horas_deporte    horas_de_trabajo
 Alto    :33    Min.   :0.0418   Min.   : 4.150
 Bajo    :35    1st Qu.:1.4059   1st Qu.: 6.663
 Moderado:29    Median :2.0295   Median : 8.286
```

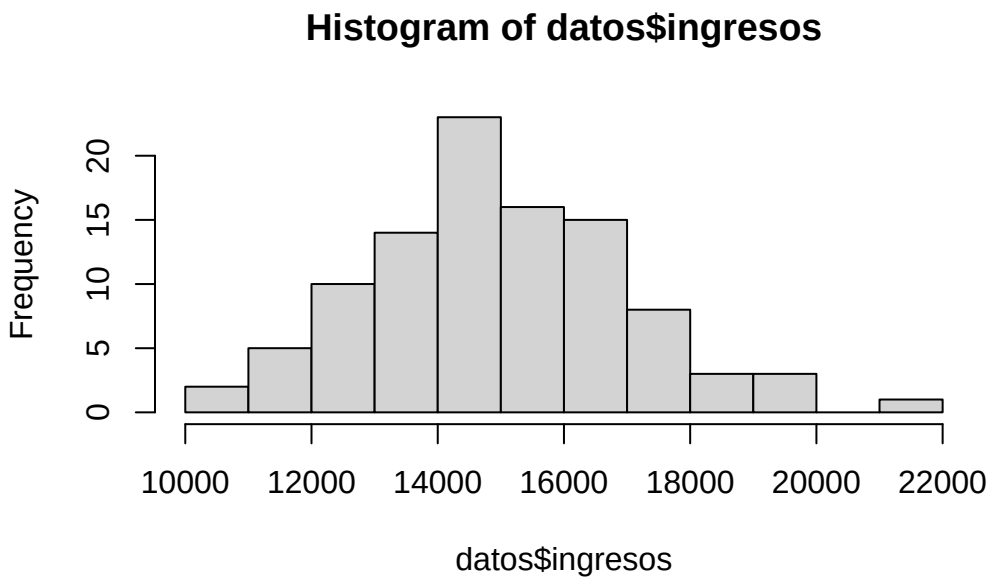
```
 NA's    : 3    Mean   :2.0380   Mean   : 8.214
                3rd Qu.:2.7541   3rd Qu.: 9.397
                Max.   :4.6849   Max.   :13.151
                NA's   :4
```

EJERCICIO 4: Realiza un gráfico de las variables cuantitativas para detectar algún posible error si no lo has hecho en el apartado anterior y contesta a la siguiente pregunta ¿Qué harías con la variable Altura?

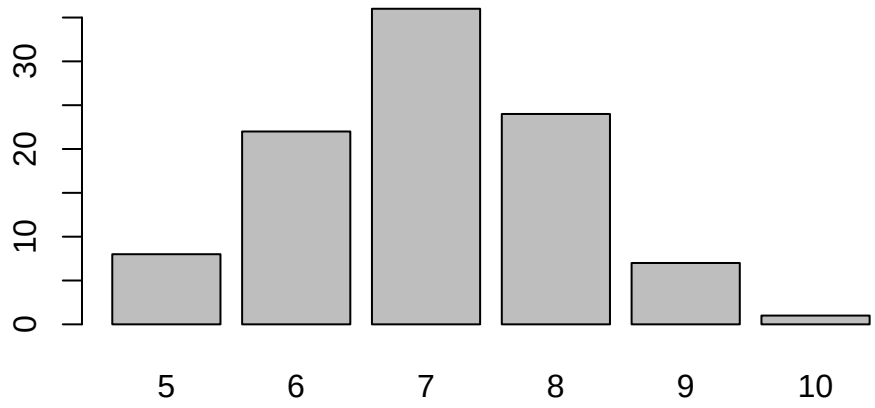
```
hist(datos$edad)
```



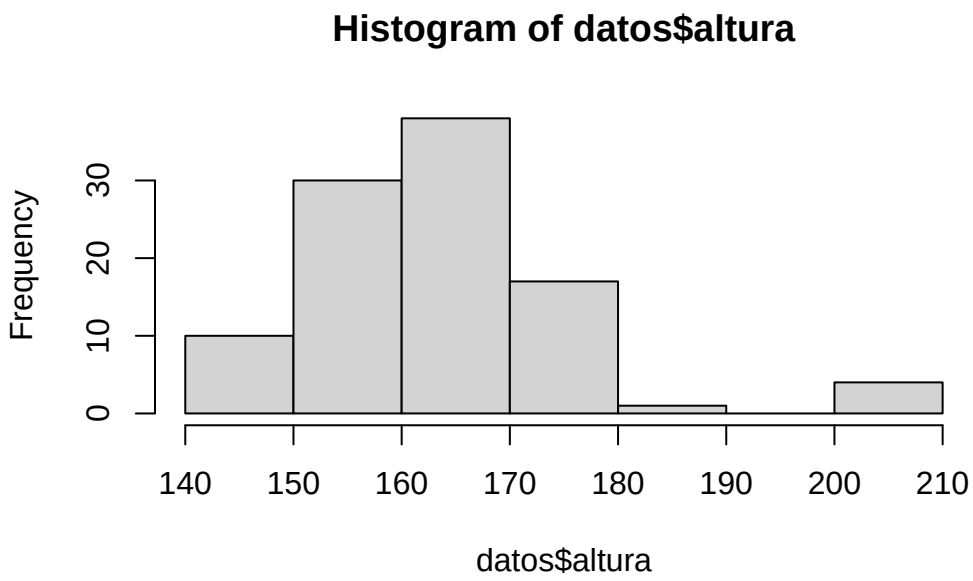

```
hist(datos$ingresos)
```



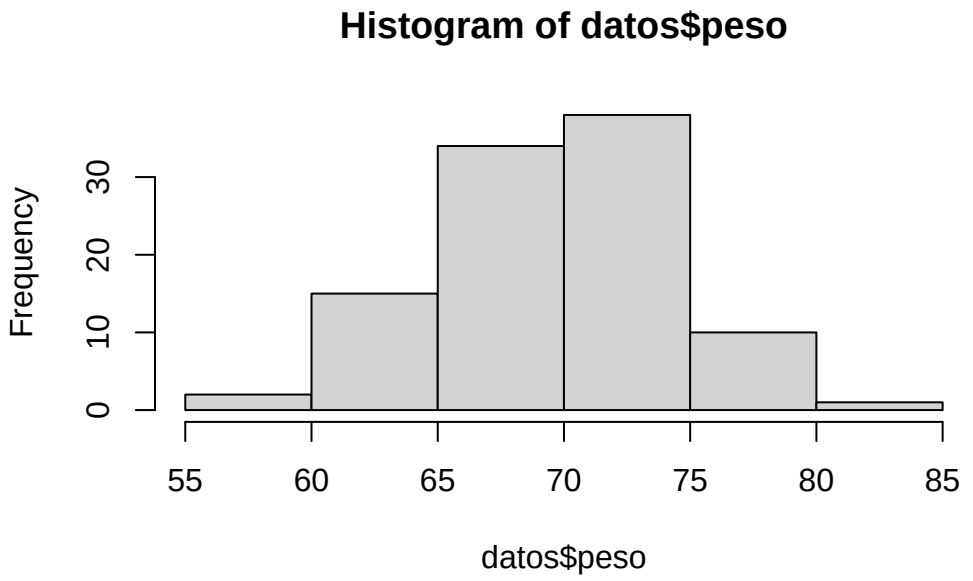

```
barplot(table(datos$puntuacion))
```



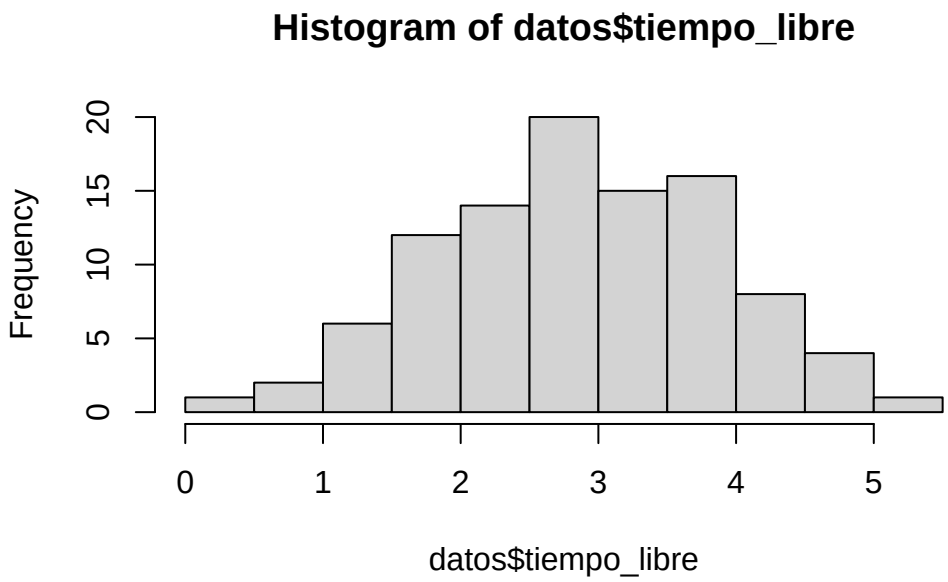
```r
hist(datos$altura)
```



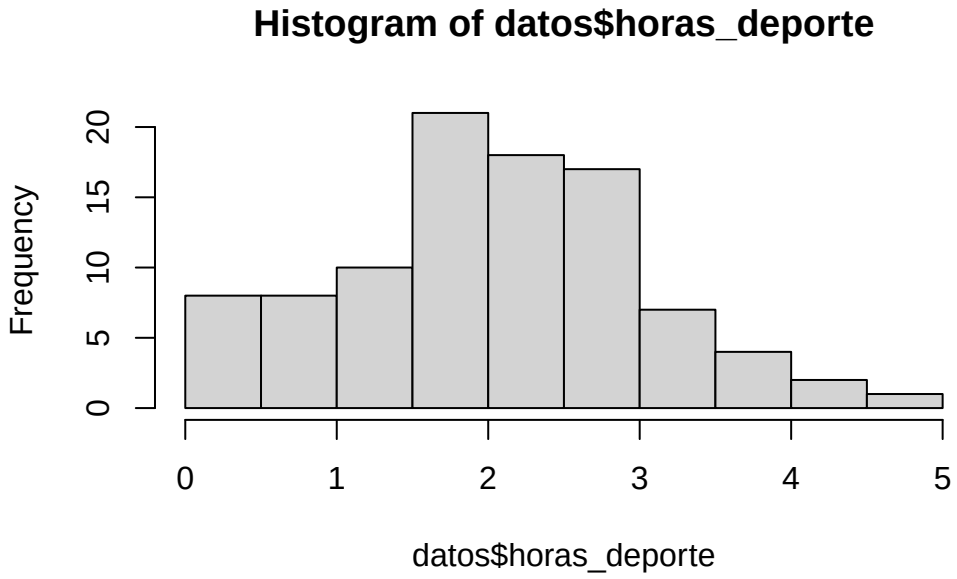

```r
hist(datos$peso)
```



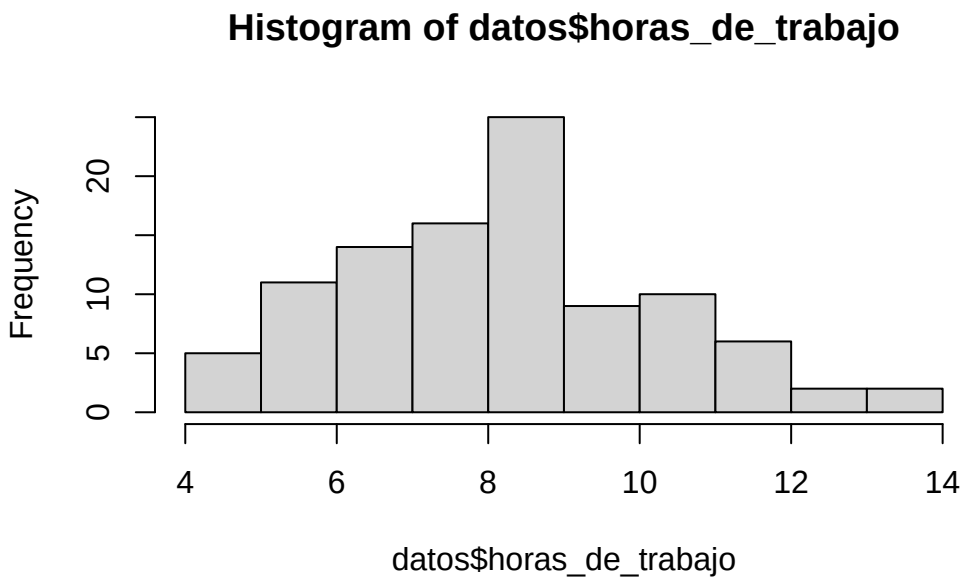
```r
hist(datos$tiempo_libre)
```



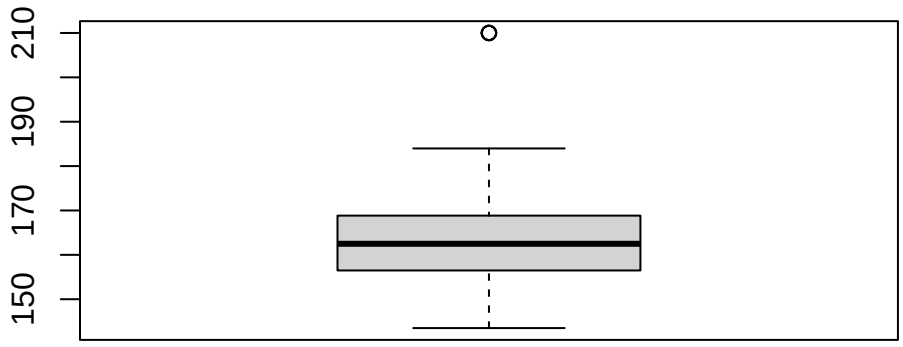

```r
hist(datos$horas_deporte)
```



```r
hist(datos$horas_de_trabajo)
```



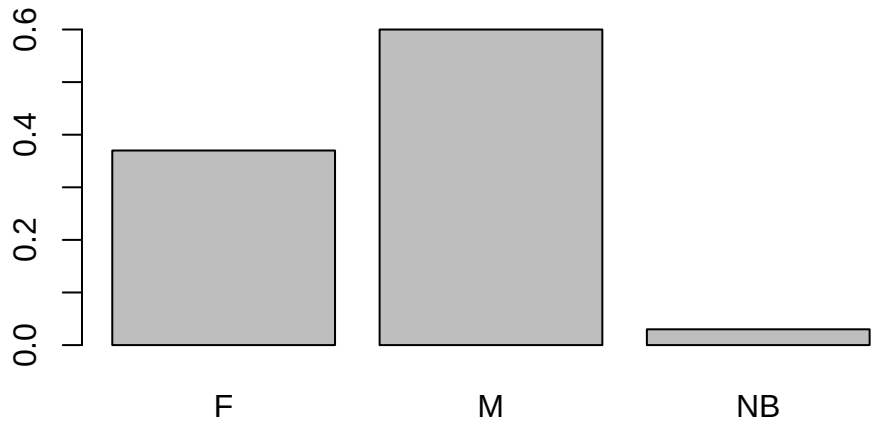

```r
boxplot(datos$altura)
```



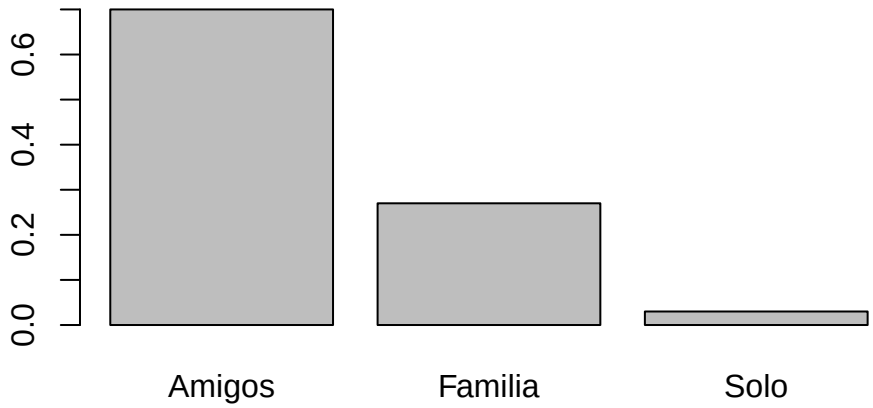
La variable altura tiene un dato extremo pero no necesariamente es un error, lo veremos en el siguiente tema.

EJERCICIO 5: Calcula las tablas de frecuencia y algún gráfico para las variables cualitativas si no lo has hecho en los apartados anteriores ¿Están equilibradas? ¿Harías algo con alguna de estas variables? Coméntalo

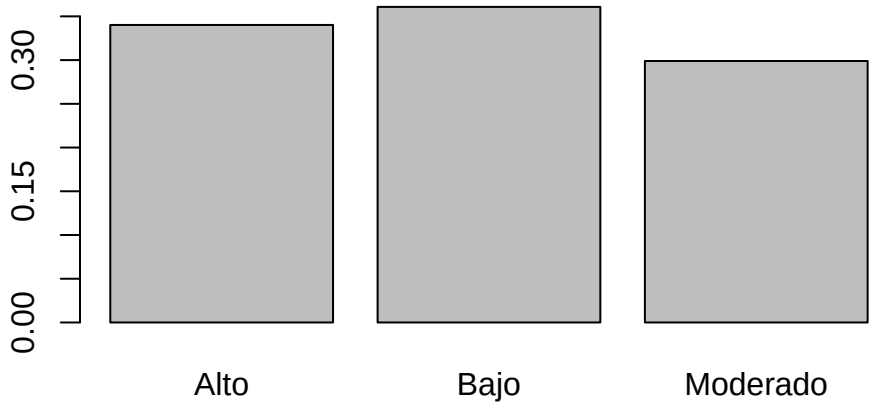
```
t1<-prop.table(table(datos$genero))
barplot(t1)
```




```
t2<-prop.table(table(datos$vivienda))
barplot(t2)
```




```
t3<-prop.table(table(datos$consumo_cafe))
barplot(t3)
```

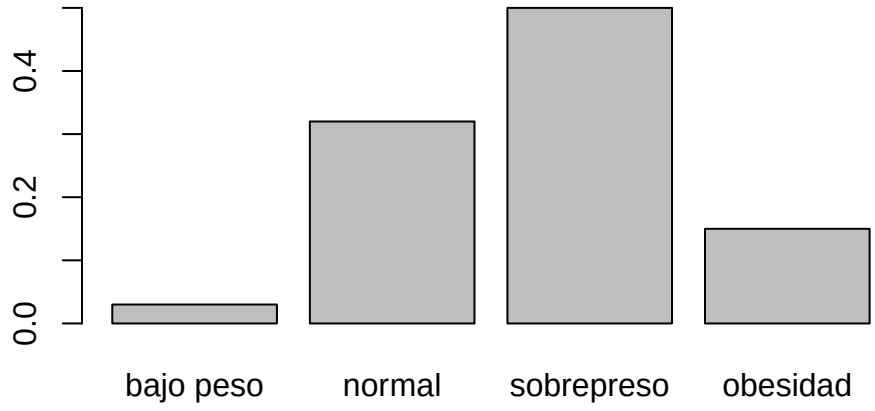


Para la variable Vivienda el número de clases de "solo" es muy pequeño y la de género la variable NB también, como todavía no hemos pensado en los análisis que vamos a realizar no hacemos nada, más que apuntarnos que hay unas categorías muy pequeñas.

EJERCICIO 6: Calcula el IMC usando la fórmula IMC=peso(kg)/altura(m)^2 y categorízalo como: bajo peso: IMC <= 18.50 normal: 18.5 < IMC <=25 sobrepeso: 25 < IMC <= 30 obesidad: IMC >30 ¿Qué porcentaje de individuos hay en cada grupo? Haz un gráfico para representarlo ¿Cuántos individuos con bajo peso viven con amigos, familia, solo? ¿Harías algo con estas variables si no lo has hecho previamente?

```
datos$IMC  <- datos$peso / (datos$altura/100)^2
datos$IMC_grupos <- cut(datos$IMC, breaks = c(-Inf,18.5,25,30,Inf),
                        labels = c("bajo peso", "normal", "sobrepreso", "obesidad"))


t<-prop.table(table(datos$IMC_grupos))
barplot(t)
```



```
t
```

```
 bajo peso     normal sobrepreso   obesidad
      0.03       0.32       0.50       0.15
```

```
datos[which(datos$IMC_grupos=="bajo peso"),]
```

```
   id edad ingresos genero vivienda puntuacion altura     peso tiempo_libre
24 24   53 15477.46      M   Amigos          8    210 66.32600    0.3983003
47 47   25 15602.31      F   Amigos          8    210 66.95721    3.6361240
64 64   25 15082.47      M  Familia          7    210 62.70515    2.1718257
   consumo_cafe horas_deporte horas_de_trabajo      IMC IMC_grupos
24     Moderado     0.8233078        10.441927 15.03991  bajo peso
47         Bajo     1.9243749         9.651846 15.18304  bajo peso
64     Moderado     2.7898179         8.730038 14.21885  bajo peso
```

Parece que la altura de 210 en 3 individuos es indicativo de muy bajo peso, es muy posible que sea un error, pero tampoco lo podemos asegurar al 100%, si lo consideramos error y lo borramos no estará mal bajo este criterio, pero cuando hay dudas de si es un error, a veces, es mejor dejarlo y ya en la siguiente fase lo excluiríamos en la detección de outliers. Además aquí tendríamos que tomar la decisión de borrar unos y no los 4.

```
prop.table(table(datos$IMC_grupos,datos$vivienda),1)
```

```
                Amigos    Familia       Solo
  bajo peso  0.66666667 0.33333333 0.00000000
  normal     0.78125000 0.21875000 0.00000000
  sobrepreso 0.62000000 0.34000000 0.04000000
  obesidad   0.80000000 0.13333333 0.06666667
```

```
datos$vivienda<-car::recode(datos$vivienda,"'Solo'= NA")
datos$IMC_grupos<-car::recode(datos$IMC_grupos,"'bajo peso'= NA")

prop.table(table(datos$IMC_grupos,datos$vivienda),1)
```

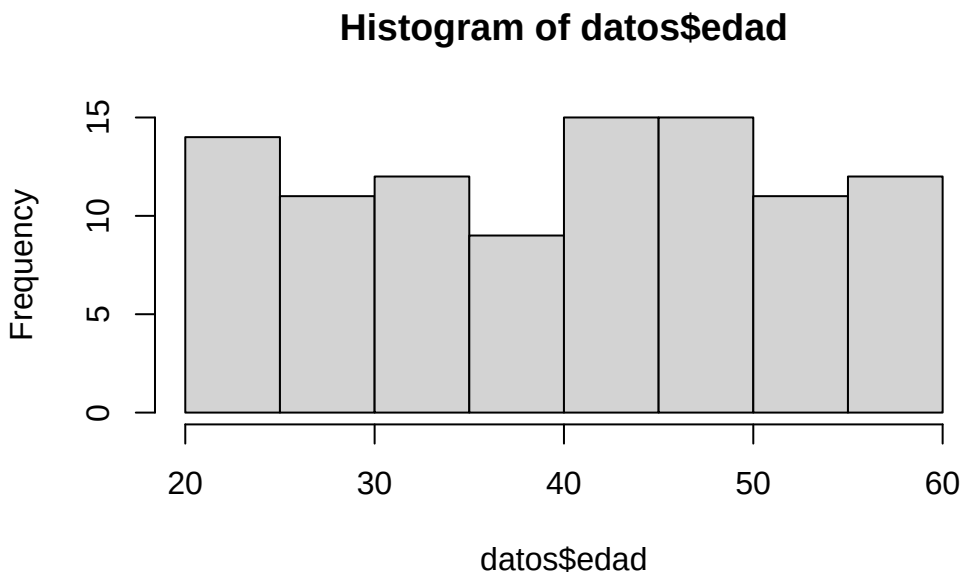
```
                Amigos   Familia
```

```
normal     0.7812500 0.2187500
obesidad   0.8571429 0.1428571
sobrepreso 0.6458333 0.3541667
```

Bajo el supuesto análisis de querer ver el porcentaje de peso según vivienda, deberíamos quitar la categoría solo y la categoría bajo peso para ver los porcenatejes más realistas

EJERCICIO 7: Queremos usar la variable edad de forma categórica, ¿Cómo la transformarías? ¿Cuántos intervalos has creado y qué porcentaje hay en cada uno de ellos?

```
hist(datos$edad)
```




```
summary(datos$edad)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  20.00   30.50   42.00   40.51   49.00   60.00       1
```

```
datos$edad_2cat<-if_else(datos$edad<=42,1,2)
table(datos$edad_2cat)
```

```
 1  2
52 47
```

```
tertiles <- quantile(datos$edad, probs = c(1/3, 2/3),na.rm = T)
```

Cuando no hay una idea clara de que hacer con la variable y la quieres catagorizar, puedes usar medianas o tertiles o cuartiles para su división

GUARDA LOS DATOS CORREGIDOS

```
write.csv(datos,"Alumnos_corregido.csv")
```