Curso programa estadístico R para el Análisis de Datos

Miguel A. Mayer

20/05/2017

Instalación de R

En primer lugar debemos instalar el programa estadístico R. Para ello debemos acceder a la web:

Instalación de RStudio

Una vez instalado R, debemos proceder a la instalación de RStudio. Para ello accedemos a la web:

RStudio

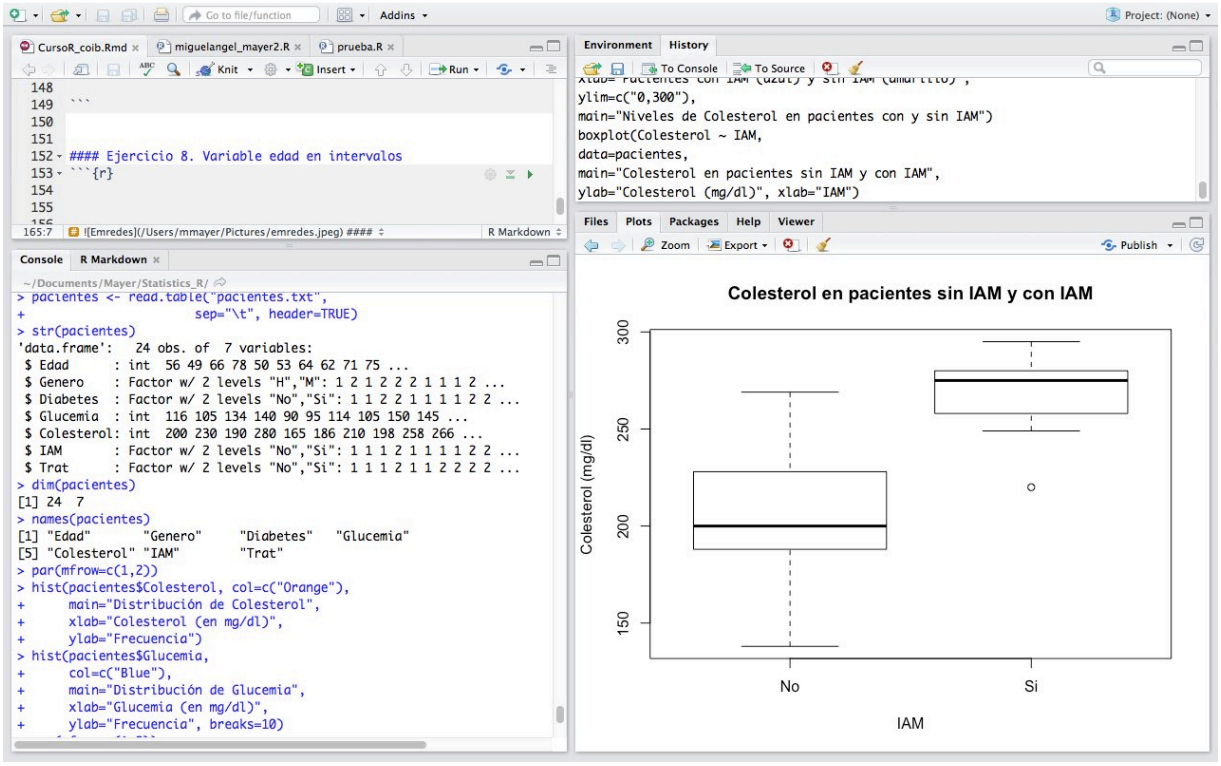


Figure 1: RStudio

Sección 1. Introducción y generalidades en estadística

Variables cualitativas:

Variables cuantitativas:

Sección 2. Primeros pasos con R y RStudio

Introducimos los datos creando dos variables: - genero - edad

```
genero <- c("h", "m", "m", "h","m","h")

edad <- c(23,45,34,39,60,52)
```

Sección 3. Importación y exportación de datos

a) Importar los datos de pacientes.txt

```
pacientes <- read.table("pacientes.txt", sep="\t", header=TRUE)
```

Sección 3. Estructura de datos

Ejercicio 2.

```
str(pacientes)
```

```
## 'data.frame':    24 obs. of  7 variables:
##  $ Edad      : int  56 49 66 78 50 53 64 62 71 75 ...
##  $ Genero    : Factor w/ 2 levels "H","M": 1 2 1 2 2 2 1 1 1 2 ...
##  $ Diabetes  : Factor w/ 2 levels "No","Si": 1 1 2 2 1 1 1 1 2 2 ...
##  $ Glucemia  : int  116 105 134 140 90 95 114 105 150 145 ...
##  $ Colesterol: int  200 230 190 280 165 186 210 198 258 266 ...
##  $ IAM       : Factor w/ 2 levels "No","Si": 1 1 1 2 1 1 1 1 2 2 ...
##  $ Trat      : Factor w/ 2 levels "No","Si": 1 1 1 2 1 1 2 2 2 2 ...
```

```
dim(pacientes)
```

```
## [1] 24  7
```

```
names(pacientes)
```

```
## [1] "Edad"       "Genero"     "Diabetes"   "Glucemia"   "Colesterol"
## [6] "IAM"        "Trat"
```

Ejercicio 3.

Ejercicio 4.

Ejercicio 5.

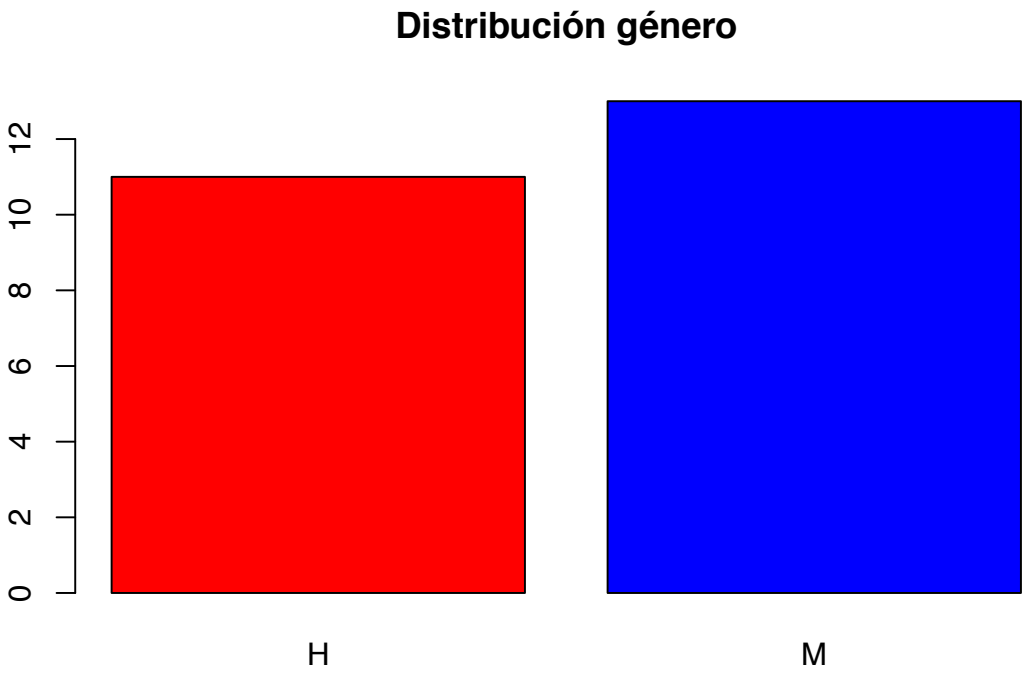
Ejercicio 6.

Sección 3. Representación gráfica

Variables cualitativas:

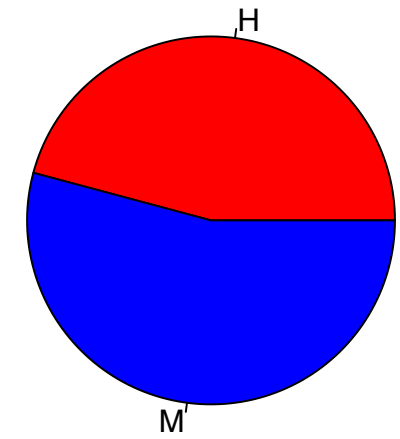
a) Diagrama de barras

```r
# Añadimos diversos parámetros para mejorar su representación
barplot(table(pacientes$Genero),
        main="Distribución género",
        col=c("red","blue"))
```



b) Diagrama de sectores

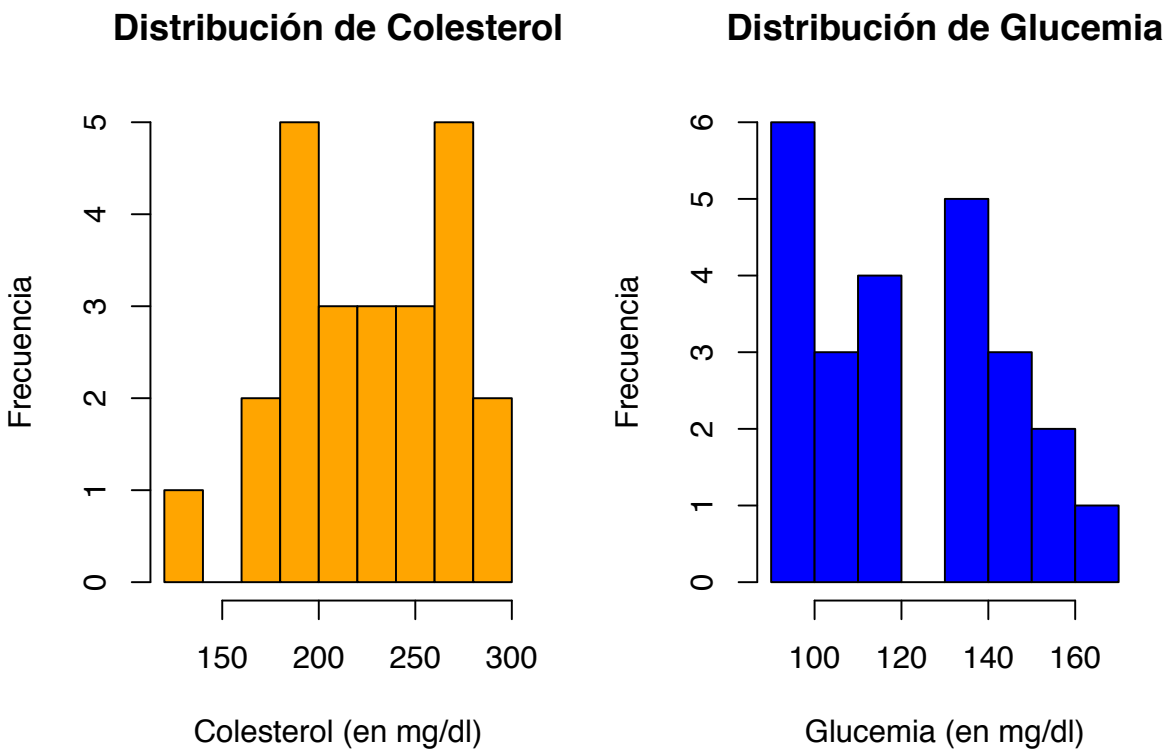
```r
pie(table(pacientes$Genero), main="Distribución por género",
                             col=c("red","blue"))
```



Variables cuantitativas:

a) Histograma

```r
par(mfrow=c(1,2))
hist(pacientes$Colesterol, col=c("Orange"),
     main="Distribución de Colesterol",
     xlab="Colesterol (en mg/dl)",
     ylab="Frecuencia")
hist(pacientes$Glucemia,
     col=c("Blue"),
     main="Distribución de Glucemia",
     xlab="Glucemia (en mg/dl)",
     ylab="Frecuencia", breaks=10)
```



b) Boxplot

Sección 4. Estadística descriptiva

Ejercicio 7.

Importamos

```r
# Tablas de datos que pueden ser intersantes para observar la frecuencia de datos
table(pacientes$Genero)
```

```
##
##  H  M
## 11 13
```

```r
table(pacientes$IAM)
```

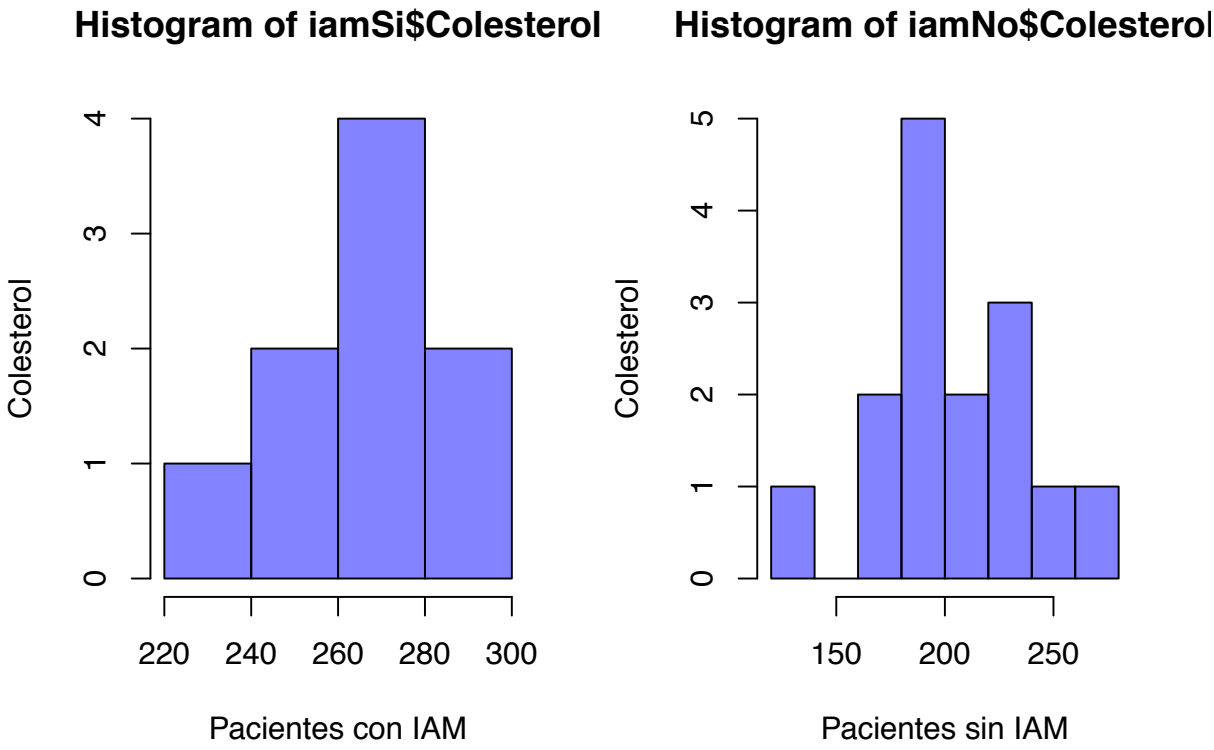
```
##
## No Si
## 15  9
```

```r
table(pacientes$Trat)
```

```
##
## No Si
## 15  9
```

Variables cuantitativas: Resumen numérico básico

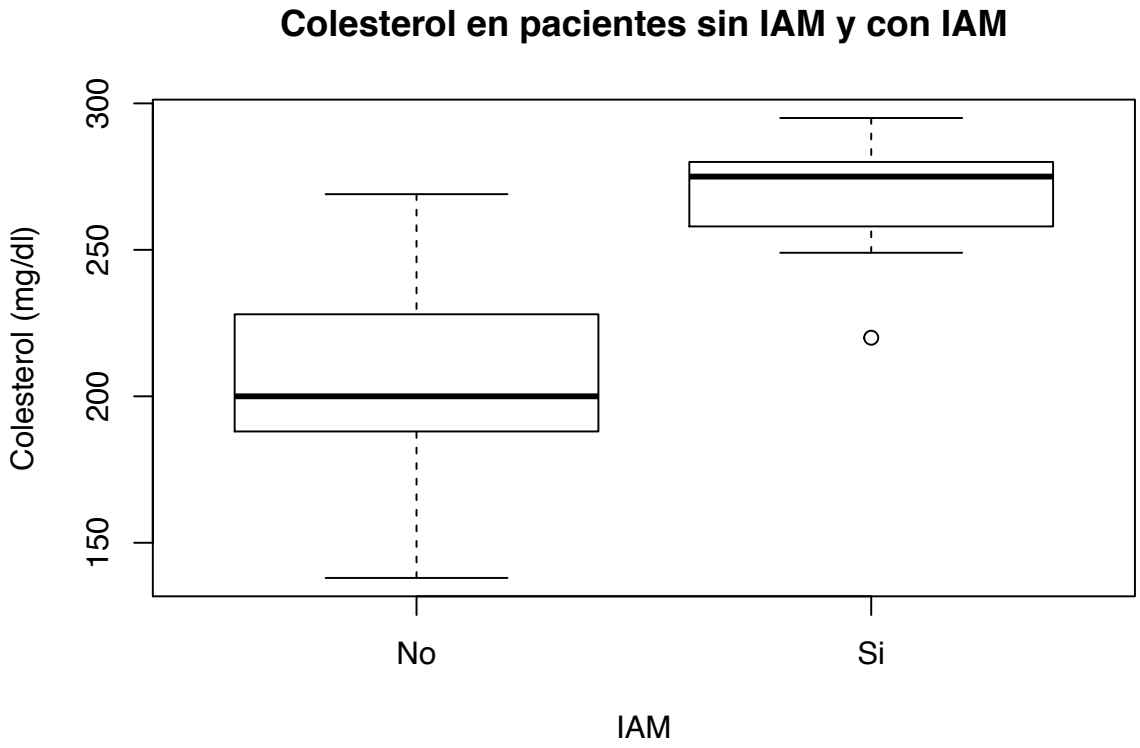
```r
# t-test para comparar las medias de colesterol en los grupos con IAM y sin IAM
# creando en primer lugar los subgrupos correspondientes
# mediante funciones como subset:
iamNo <- subset(pacientes, IAM=="No", select="Colesterol") # o selec=c(Colesterol)
iamSi <- subset(pacientes, IAM=="Si", select="Colesterol")

# otra forma de realizar subsetting: con "which"
# iamNo <- pacientes[which(pacientes$IAM == "No"), "Colesterol"]
# iamSi <- pacientes[which(pacientes$IAM == "Si"), "Colesterol"]

###### representación para comparar los valores de colesterol entre pacientes
###### con IAM y sin IAM
# creación de colores transparentes
mycol2 <- rgb(255,255,0, max=255, alpha=125) # color amarillo
mycol <- rgb(0,0,255, max=255, alpha=125) # color azul
# gráficos paralelos
par(mfrow=c(1,2))
hist(iamSi$Colesterol,
        col=mycol,
        xlab="Pacientes con IAM",
        breaks=5,
        ylab="Colesterol")
hist(iamNo$Colesterol,
     col=mycol,
     xlab="Pacientes sin IAM",
     ylab="Colesterol")
```



Histogramas

Boxplots

```
# comparación mediante boxplot
boxplot(Colesterol ~ IAM,
        data=pacientes,
        main="Colesterol en pacientes sin IAM y con IAM",
        ylab="Colesterol (mg/dl)", xlab="IAM")
```



t-test

```
# realización del t-test
t.test(iamNo, iamSi, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  iamNo and iamSi
## t = -4.8355, df = 22, p-value = 7.853e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -88.65450 -35.43439
## sample estimates:
## mean of x mean of y
##  206.0667  268.1111
```

Obtenemos un valor de $p < 0.05$ y por tanto existen diferencias significativas entre ambos grupos de pacientes en cuanto a la presencia de IAM y colesterol, siendo la media de 206,06 mg/dl en aquellos que no tienen IAM y de 268,11 en los que presentan IAM

Regresión simple

Regresión simple entre edad y glucemia y colesterol siendo la Edad la variable independiente.

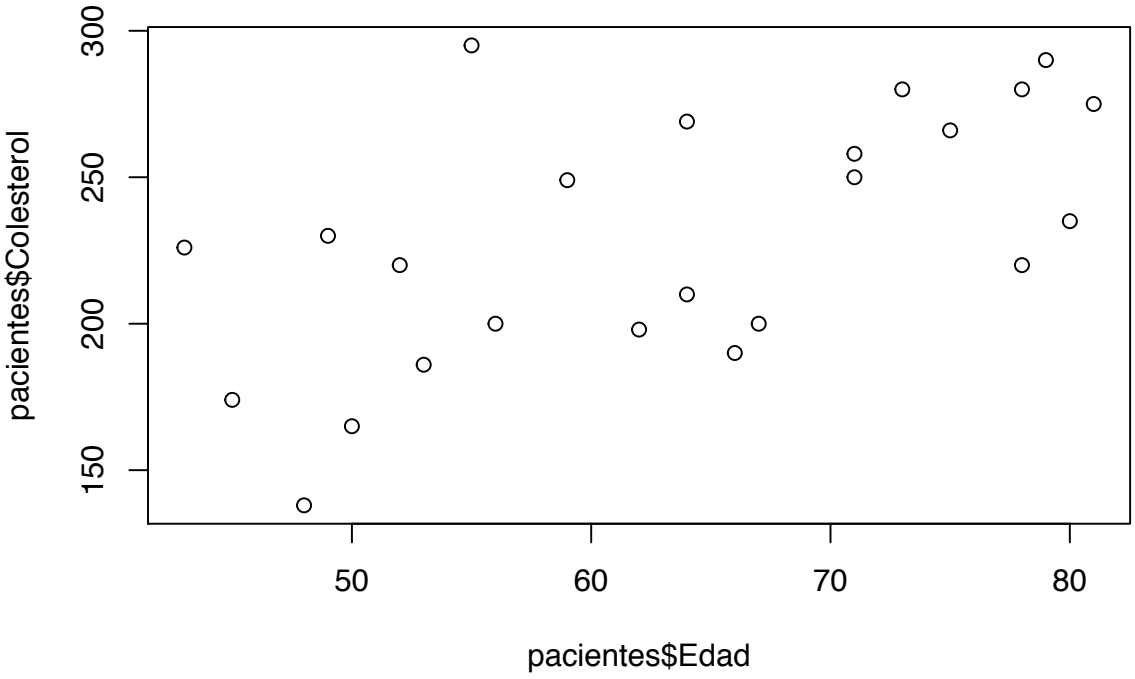
```
##### Regresión simple entre edad y glucemia y colesterol.
# siendo la Edad la variable independiente
fit <- lm(pacientes$Glucemia ~ pacientes$Edad)
summary(fit)
```

```
##
## Call:
## lm(formula = pacientes$Glucemia ~ pacientes$Edad)
```

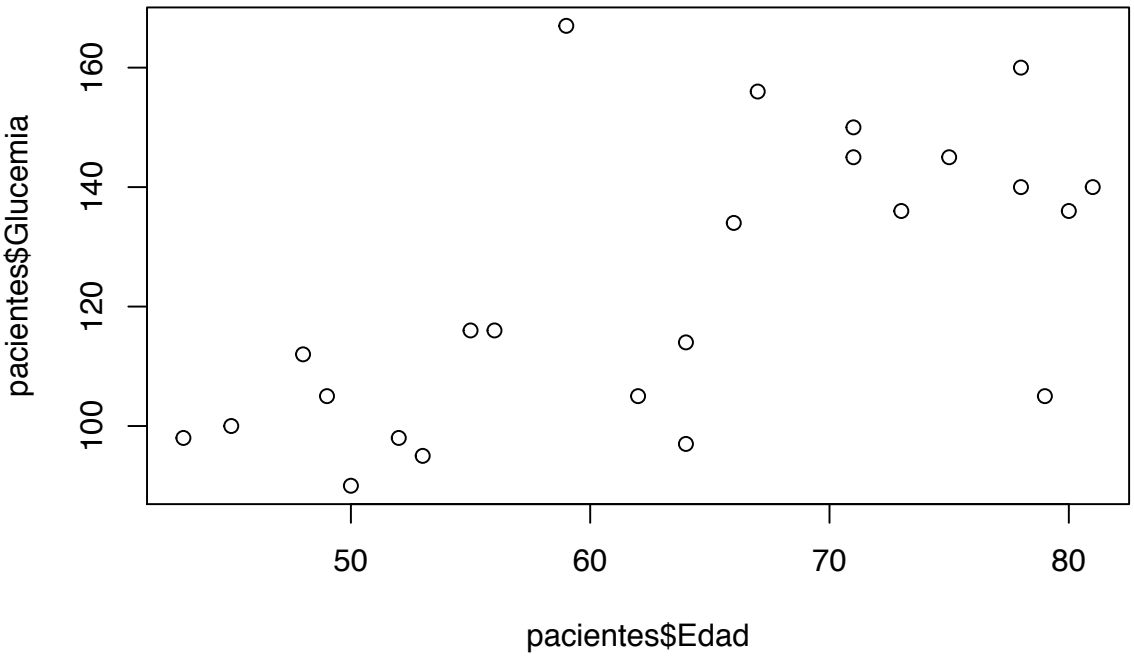
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.962 -10.470  -0.227   7.405  49.029
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     44.2474    20.3748   2.172 0.040939 *
## pacientes$Edad   1.2495     0.3164   3.949 0.000683 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.4 on 22 degrees of freedom
## Multiple R-squared:  0.4148, Adjusted R-squared:  0.3882
## F-statistic:  15.6 on 1 and 22 DF,  p-value: 0.0006826
```

```
# Nos indica que con una p < 0.05 la edad influye en los valores de
# glucemia y la recta de regresión es: glucemia = 44.247 + 1.2495*edad
# A mayor edad, valores de glucemia más altos
```

```
##### Gráficos de ambas variables
plot(pacientes$Colesterol ~ pacientes$Edad)
```




```
plot(pacientes$Glucemia ~ pacientes$Edad)
```



Ejercicio 8. Variable edad en intervalos

Transformación de la variable edad en determinados intervalos.