Taller Introductorio de AED

Taller diseñado por : Santiago Alférez

Enero de 2022

A continuación veremos algunos tipos de gráficas para realizar análisis exploratorio de datos (EDA) a un conjunto de observaciones.

Instalación de librerias para los ejemplos

```
# Si un paquete es instalado, entonces solo sera cargado. Si
# hay paquetes no instalados, entonces serán instalados desde
# CRAN y serán cargados

## Paquetes de interes
packages = c("dslabs", "MASS", "scatterplot3d", "car")

## Se cargan o se instalan y cargan
package.check <- lapply(
  packages,
  FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
      install.packages(x, dependencies = TRUE)
      library(x, character.only = TRUE)
    }
  }
)
```

```
## Loading required package: dslabs
## Warning: package 'dslabs' was built under R version 3.6.3
## Loading required package: MASS
## Warning: package 'MASS' was built under R version 3.6.3
## Loading required package: scatterplot3d
## Loading required package: car
## Loading required package: carData
```

Descripción del dataset `Olive`

It contains 572 rows of observations. The first and the second column correspond to the area (Centre-North, South, Sardinia) and the geographical region of origin of the olive oils (northern Apulia, southern Apulia, Calabria, Sicily, inland Sardinia and coast Sardinia, eastern and western Liguria, Umbria), respectively. The remaining columns represent the chemical measurements (on the acid components for the oil specimens) palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic.

```r
head(olive)
```

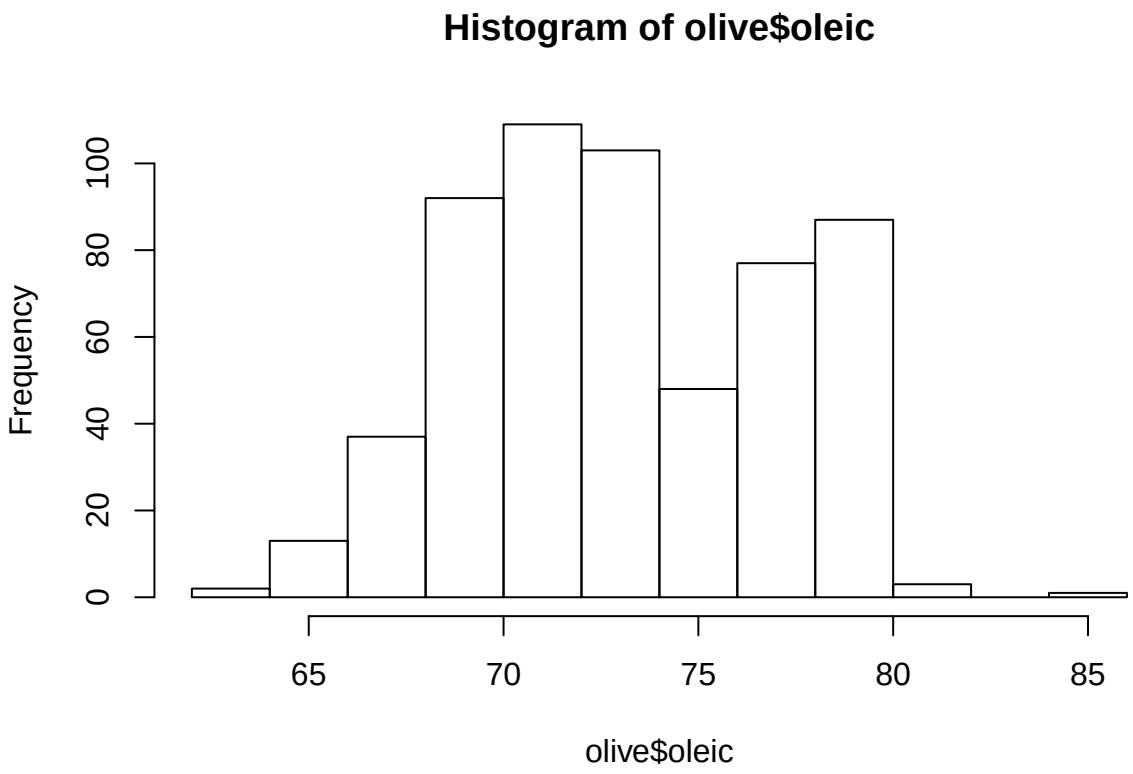
```
##           region         area palmitic palmitoleic stearic oleic linoleic
## 1 Southern Italy North-Apulia    10.75        0.75    2.26 78.23     6.72
## 2 Southern Italy North-Apulia    10.88        0.73    2.24 77.09     7.81
## 3 Southern Italy North-Apulia     9.11        0.54    2.46 81.13     5.49
## 4 Southern Italy North-Apulia     9.66        0.57    2.40 79.52     6.19
## 5 Southern Italy North-Apulia    10.51        0.67    2.59 77.71     6.72
## 6 Southern Italy North-Apulia     9.11        0.49    2.68 79.24     6.78
##   linolenic arachidic eicosenoic
## 1      0.36      0.60       0.29
## 2      0.31      0.61       0.29
## 3      0.31      0.63       0.29
## 4      0.50      0.78       0.35
## 5      0.50      0.80       0.46
## 6      0.51      0.70       0.44
```

Gráficas de una sola variable

Histograma

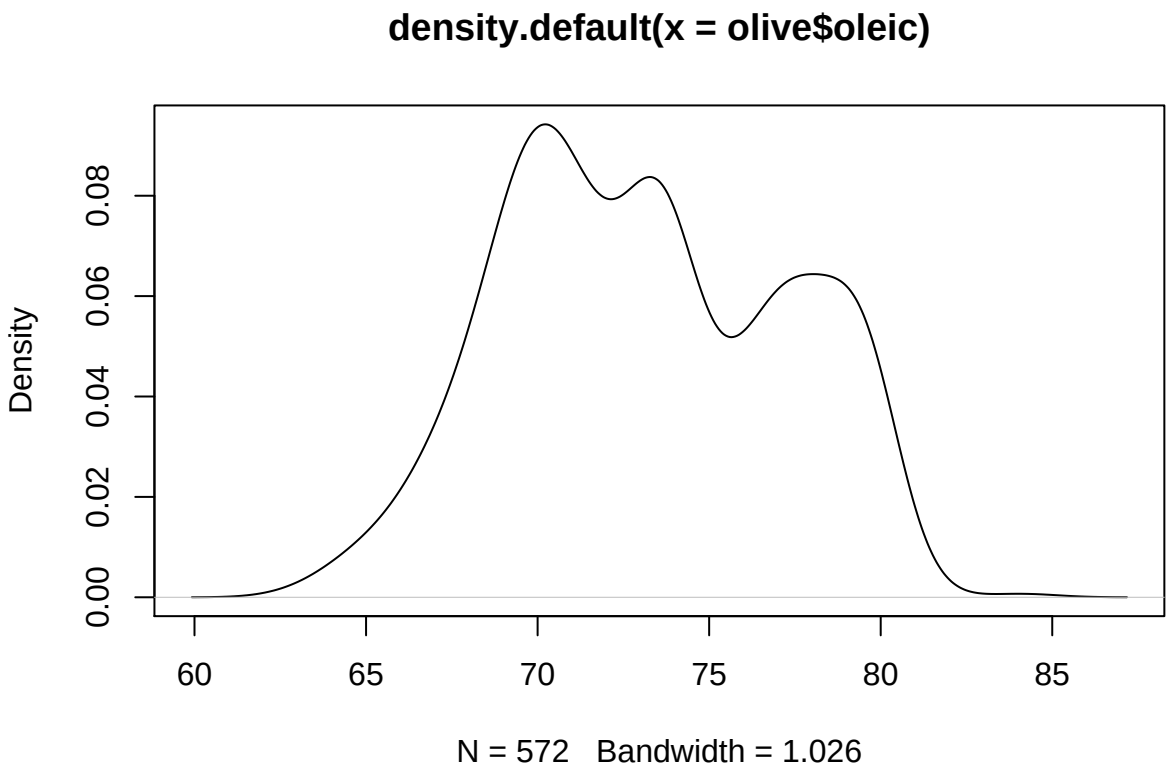
Es la gráfica adecuada para representar variables cuantitativas con un gran número de valores distintos. Los datos se agrupan en intervalos y se representan gráficamente por rectángulos yuxtapuestos cuyas bases descansan sobre el eje horizontal y cuyas alturas son tales que el área de cada rectángulo sea proporcional a la frecuencia de cada intervalo. Si todos los intervalos tienen igual longitud, entonces la altura de cada rectángulo es proporcional a la frecuencia el intervalo. Para evitar confusiones, la principal diferencia con el gráfico de barras es la inexistencia de espacios entre rectángulos. La función `hist()` permite hacer el histograma de unos datos y además modificar la longitud de los intervalos si se desea. A diferencia del gráfico de barras, la función calcula automáticamente la frecuencia del intervalo.

```r
hist(olive$oleic)
```



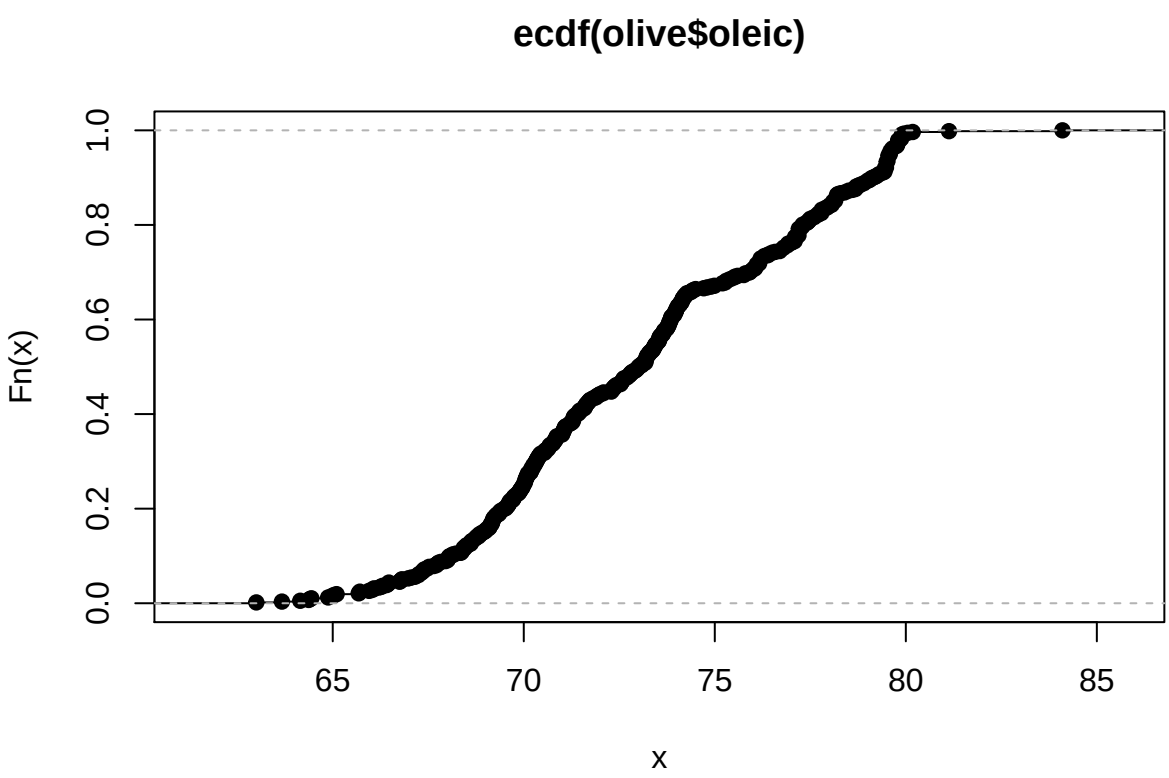
Gráficas de densidad

```
plot(density(olive$oleic))
```



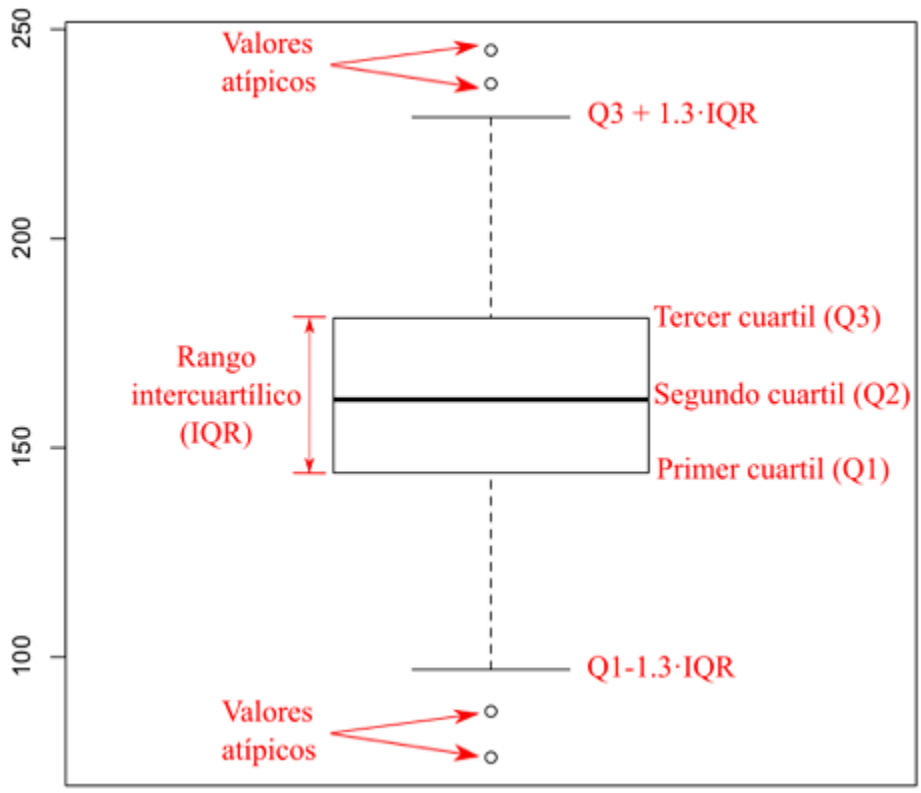
ECDF (Función de distribución empírica)

```
plot(ecdf(olive$oleic))
```



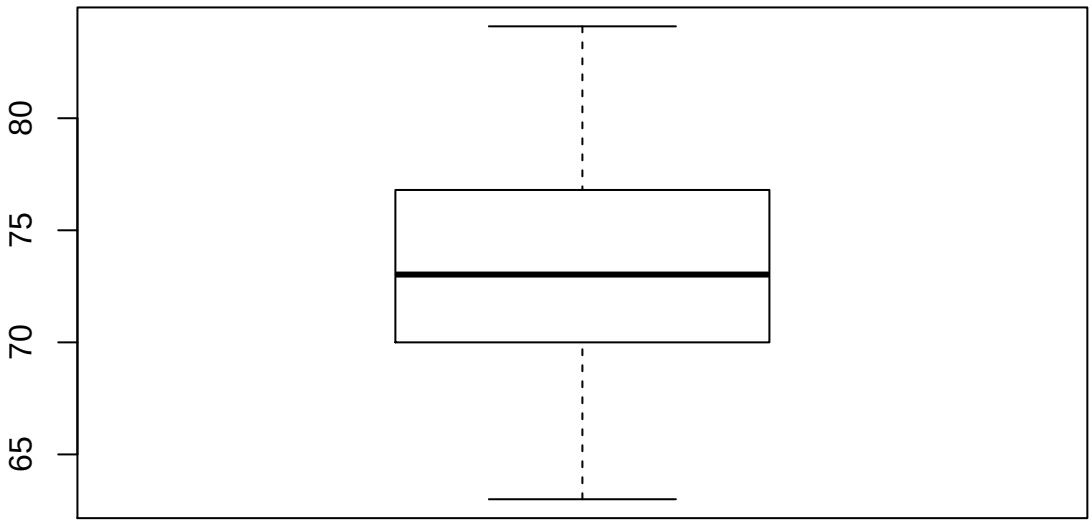
Boxplots

Los diagramas de caja son una presentación visual que describe varias características importantes al mismo tiempo, tales como la tendencia central, dispersión y simetría. Para su realización se representan los tres cuartiles y los valores mínimo y máximo de los datos sobre un rectángulo, alineado horizontal o verticalmente. Los valores con dispersión hasta 1.3 (o 1.5) veces el rango intercuartílico se representan como unas líneas rectas o bigotes. Los valores fuera de ese intervalo se representan mediante puntos y se consideran valores extremos atípicos.



`boxplot()` es la función que se utiliza para la creación del gráfico. Tal cual como con el histograma, si se guarda el gráfico de caja en un objeto `h = boxplot()`, éste objeto contiene información como los límites para considerar los valores atípicos, cuáles son esos valores atípicos, los cuartiles, etc.

```
boxplot(olive$oleic)
```



Descripción del dataset `stars`

Physical properties of selected stars, including luminosity, temperature, and spectral class.

star: Name of star. magnitude: Absolute magnitude of the star, which is a function of the star's luminosity and distance to the star. temp: Surface temperature in degrees Kelvin (K). type: Spectral class of star in the OBAFGKM system.

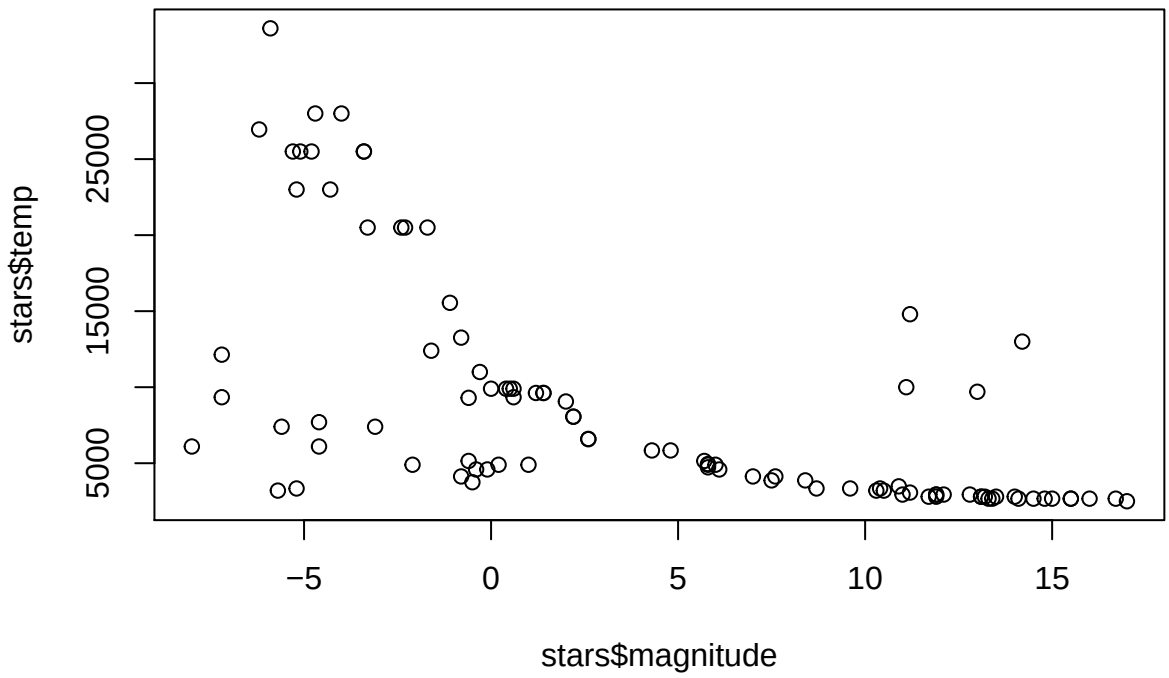
```
head(stars)
```

```
##             star magnitude temp type
## 1            Sun       4.8 5840    G
## 2        SiriusA       1.4 9620    A
## 3        Canopus      -3.1 7400    F
## 4       Arcturus      -0.4 4590    K
## 5 AlphaCentauriA       4.3 5840    G
## 6           Vega       0.5 9900    A
```

Gráficas de dos variables

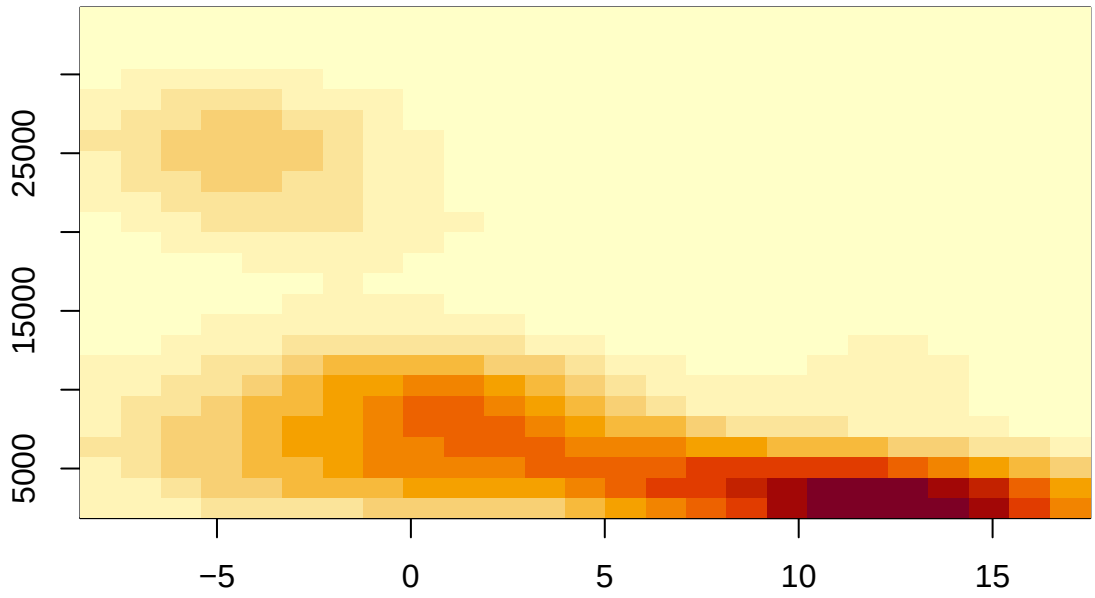
Diagrama de dispersión

```
plot(stars$magnitude,
     stars$temp)
```



Gráficas de densidad (2D)

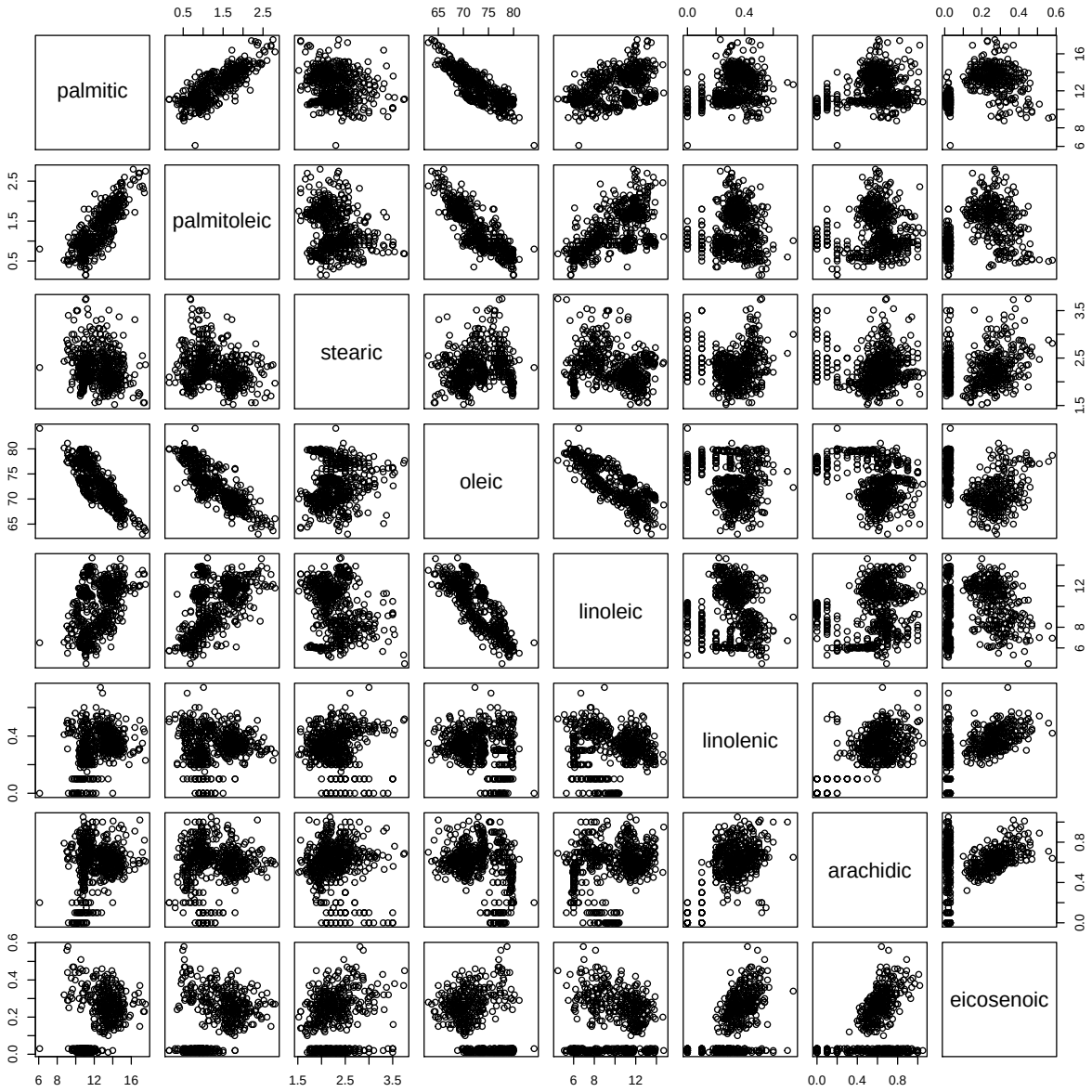
```
image(kde2d(stars$magnitude,
            stars$temp))
```



Pairplots

Son matrices de gráficas, es decir una composición cuadrada de gráficas (por ejemplo de dispersión) en donde cambian las variables dependiendo de la posición de la matriz. Una forma sencilla es graficando directamente el data.frame (o una parte) con la función `plot`

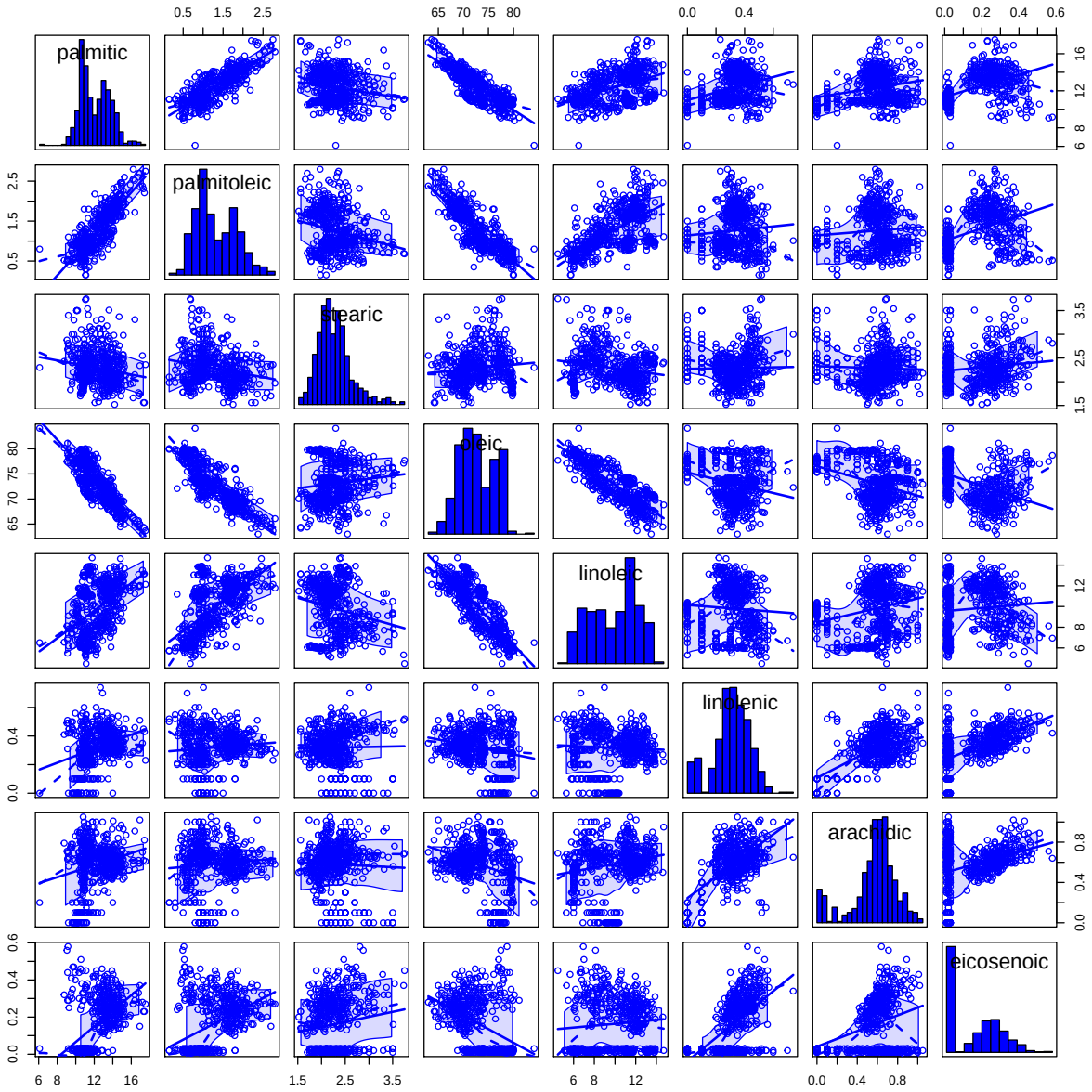
```
olive_acids <- subset(olive, select = c(-region, -area))
plot(olive_acids)
```



Pairplot usando la libreria `car`

R tiene muchísimas librerías que pueden ayudar a graficar o a implimentar métodos (como veremos a lo largo del curso). Algunas librerías están en el R base (sólo necesitan ser cargadas), otras hay que instalarlas (como al inicio de este notebook). Particularmente, vamos a usar la librería `car` para graficar muy fácilmente una matriz de dispersión similar a la anterior.

```
scatterplotMatrix(~ +., data = olive_acids, diagonal=list(method ="histogram"))
```



Gráfica en 3D

Diagrama de dispersión 3D

Es importante recordar cómo se lee un archivo de datos y que objeto lo representa en R.

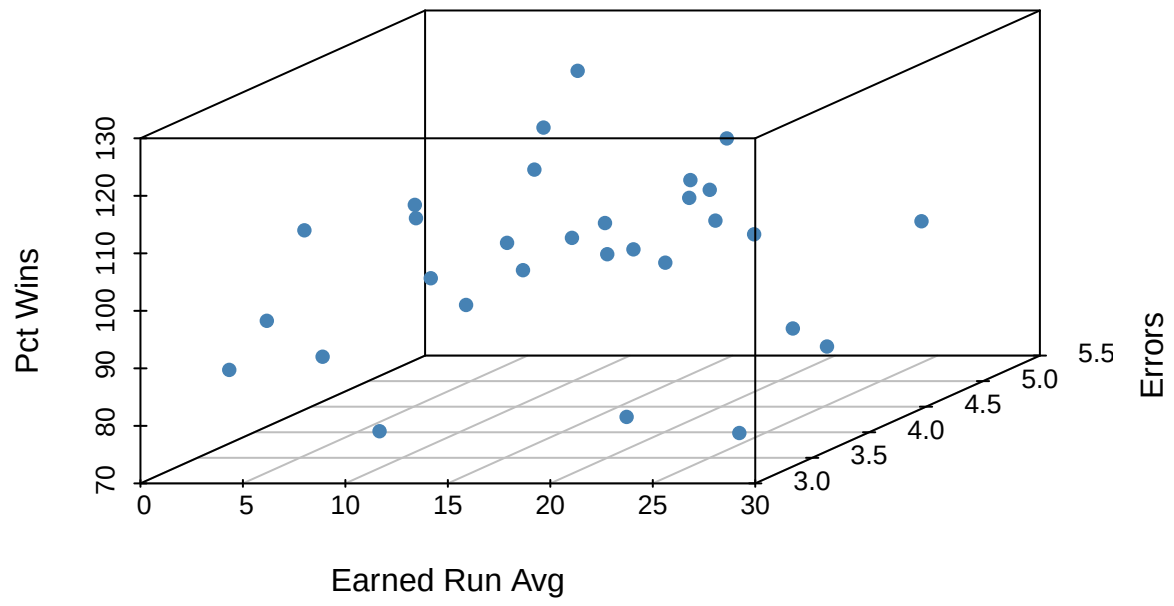
```
MLB<-read.csv("Major League Baseball Main Stats.csv")
head(MLB)
```

```
##                   Team Mean...Salary Pct..Wins Batting.Avg. Earned.Run.Avg.
## 1 Arizona Diamondbacks     2,653,029     0.500        0.259            3.93
## 2       Atlanta Braves     2,776,998     0.580        0.247            3.42
## 3    Baltimore Orioles     2,807,896     0.574        0.247            3.90
## 4       Boston Red Sox     5,093,724     0.426        0.260            4.70
## 5         Chicago Cubs     3,392,193     0.377        0.240            4.51
```

```
## 6    Chicago White Sox     3,876,780      0.525         0.255             4.02
##   Errors
## 1     90
## 2     86
## 3    106
## 4    101
## 5    105
## 6     70
```

Grafica un diagrama de dispersión en 3D es algo más complicado, pero podemos hacerlo a través de la librería de `scatterplot3d`:

```
scatterplot3d(MLB$Earned.Run.Avg,MLB$Errors,MLB$Pct.Wins,pch = 16, color="steelblue",
              xlab="Earned Run Avg", ylab="Errors", zlab="Pct Wins")
```



Ejercicios para entregar

Estos ejercicios se deben entregar en un reporte realizado en Rmarkdown (notebook). Se debe entregar tanto el archivo `.Rmd` como el archivo `.html` del notebook (no olvidar los nombres). El reporte debe estar organizado, de tal forma que después del enunciado sigue la solución. Favor borrar el contenido explicativo anterior o crear un nuevo notebook Rmardown.

1. Graficar los boxplots de la variable oleic vs region del dataset olive.
2. Los siguientes datos se extrajeron de la revista Motor Trend 1974 de Estados Unidos, resume el consumo y 10 aspectos de diseño y rendimiento de 32 automóviles (modelos 1973-74). Este conjunto de datos, que se llama mtcars, contiene 11 variables con 32 observaciones y está almacenado en R . Para poder

trabajar con ellos, solo hace falta adjudicarle un nombre al objeto, como por ejemplo:

```
a = mtcars
head(a)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Las variables son las siguientes:

- mpg: Millas por galón de combustible
- cyl: Número de cilindros
- disp: Desplazamiento
- hp: Caballos de potencia
- drat: Relación del eje trasero
- wt: Peso (1000 lbs)
- qsec: Tiempo a 1/4 milla
- vs: V/S
- am: Transmisión (0 = automático, 1 = manual)
- gear: Número de marchas adelante
- carb: Número de carburadores

Una vez cargado el conjunto de datos proceda a la resolución del cuestionario, añadiendo los *chunks* correspondientes.

- Determine la media, la mediana, la moda y la desviación estándar de cada una de las variables. Se puede calcular a todas la variables? a cuales no? Justifique su respuesta
- Determinar qué variable presenta valores atípicos, ¿cómo los ha encontrado?
- Hacer el histograma para cada una de las variable usando 5 intervalos. De nuevo, está gráfica es útil para todas las variables? justifique su respuesta.
- Realice una gráfica que incluya el diagrama de cajas de todas las variables de tal manera de que se puedan comparar.

3. Graficar una matrix de dispersion de tres variables del dataset olive, con la diagonal mostrando boxplots de las variables.

4. Graficar un diagrama de dispersión en 3D, de tres variables numéricas del dataset `olive` graficando en colores diferentes las regiones.

5. Dados los siguientes pares de medidas sobre dos variables x_1 y x_2:

x_1	-6	-3	-2	1	2	5	6	8
x_2	-2	-3	1	-1	2	1	5	3

(a) Grafique los datos como un diagrama de dispersión y calcule s_{11}, s_{22} y s_{12}.

(b) Usando $\tilde{x}_1 = x_1\cos(\theta) + x_2\sin(\theta)$ y $\tilde{x}_2 = -x_1\sin(\theta) + x_2\cos(\theta)$, calcule las medidas correspondientes sobre las variables $\widetilde{x}_1$ y $\widetilde{x}_2$, asumiendo que los ejes coordenados originales están rotados un ángulo de $\theta = 26$ grados.

(c) Usando las medidas $\tilde{x}_1$ y $\tilde{x}_2$ de (b), calcule las varianzas de muestra $\widetilde{s}_{11}$ y $\widetilde{s}_{22}$

(d) Considere el nuevo par de medidas $(x_1, x_2) = (4, -2)$. Transforme estas medidas en $\tilde{x}_1$ y $\tilde{x}_2$ como en (b) y calcule la distancia $d(O, P)$ del nuevo punto $P = (\widetilde{x}_1, \tilde{x}_2)$ desde el origen $O = (0, 0)$, usando $d(O, P) = \sqrt{\frac{\widetilde{x}_1^2}{\widetilde{s}_{11}} + \frac{\widetilde{x}_2^2}{s_{22}}}$. Nota: Necesitará $\widetilde{s}_{11}$ y $\widetilde{s}_{22}$ de (c).

(e) Calcule la distancia desde $P = (4, -2)$ hasta el origen $O = (0, 0)$ usando $d(O, P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$ y las expresiones para a_{11}, a_{22}, y a_{12} de la página 35 del libro. Nota: necesitará s_{11}, s_{22}, y s_{12} de (a). Compare la distancia calculada aquí con la distancia calculada usando los valores $\widetilde{x}_1$ y $\widetilde{x}_2$ en (d). (Dentro del error de redondeo, los números deben ser los mismos).

6. Sea $\mathbf{x}' = (x_1, x_2)$ un vector de $\mathbb{R}^2$: de Determine si las siguientes funciones corresponden a distancias válidas de $\mathbb{R}^2$:

a. $x_1^2 + 4x_2^2 + x_1x_2 = d(0, \mathbf{x})^2$
b. $x_1^2 - 2x_2^2 = d(0, \mathbf{x})^2$

7. Muestre que la matriz de correlación muestral también se puede obtener como la matrix de covarianza de las observaciones estandarizadas.