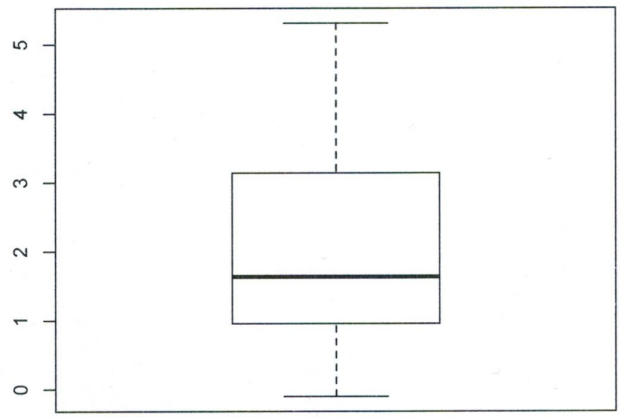
```
>boxplot(x, axes=TRUE, frame=TRUE)
```

Figura B datos simulados



Hasta el momento, hemos ilustrado que la posible existencia de subgrupos en un conjunto de datos, puede ser detectado (al menos en el caso unidimensional) usando histogramas y estimaciones de la densidad. Recomendamos al lector que revise estos temas en libros de Análisis Estadístico de Datos y Analysis de datos Multivariados, por ejemplo: "Data Analysis and Graphics Using R: An example-based approach", John Maindonald

and John Braun, Cambridge University Press.

De acuerdo a lo que hemos discutido, una opción para determinar si alguna componente de $\mathbb{X} = (X_1, X_2, \ldots, X_6)$ nos ayuda a encontrar subgrupos en los datos, sería producir una gráfica con histogramas y/ó densidades estimadas para todas las componentes en $\mathbb{X}$. No obstante, vamos a aprovechar para presentar otra gráfica de datos que suele usarse para estudiar comportamientos y posibles dependencias de datos con dimensión mayor a 1.

DIAGRAMAS DE DISPERSION (Scatterplots)

Los diagramas de dispersión son gráficos de varias componentes del vector $\mathbb{X} = (X_1, \ldots, X_n)$.

Por ejemplo, para el caso de los billetes del banco Suizo $\mathbb{X} = (X_1, \ldots, X_6)$ y

dados los datos $x_{G1}, \ldots, x_{G200}$, donde $x_{Gi} = (x_{i1}, \ldots, x_{i6})$, podemos graficar los puntos (x_{i1}, x_{i2}); $i = 1, 2, \ldots, 200$ en $\mathbb{R}^2$, lo cual nos da idea de cómo varía el vector (X_1, X_2)

```
> datos <- read.table("SwissBank 1.txt")
> x <- datos[,1]
> y <- datos[,2]
> plot(x,y)
```

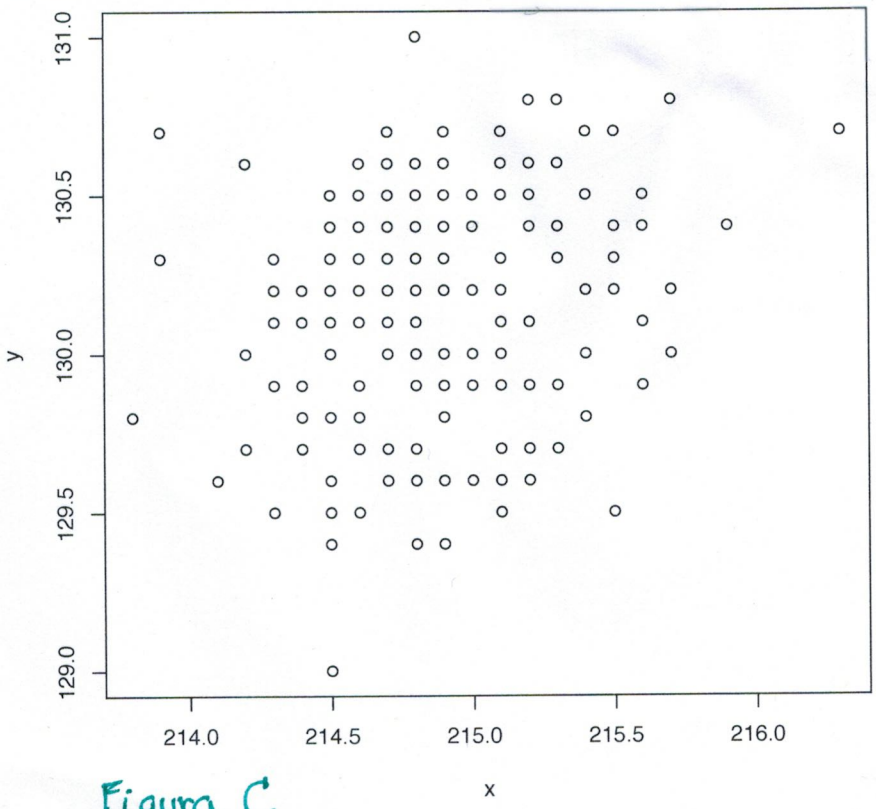


Figura G

Se pueden explorar todas las posibilidades, i.e. graficar (x_{ik}, x_{il}) ; $i=1,2,..,200$; $k \neq l$ $k,l \in \{1,2,..,6\}$. La buena noticia es que ya existen funciones en R para hacer esto

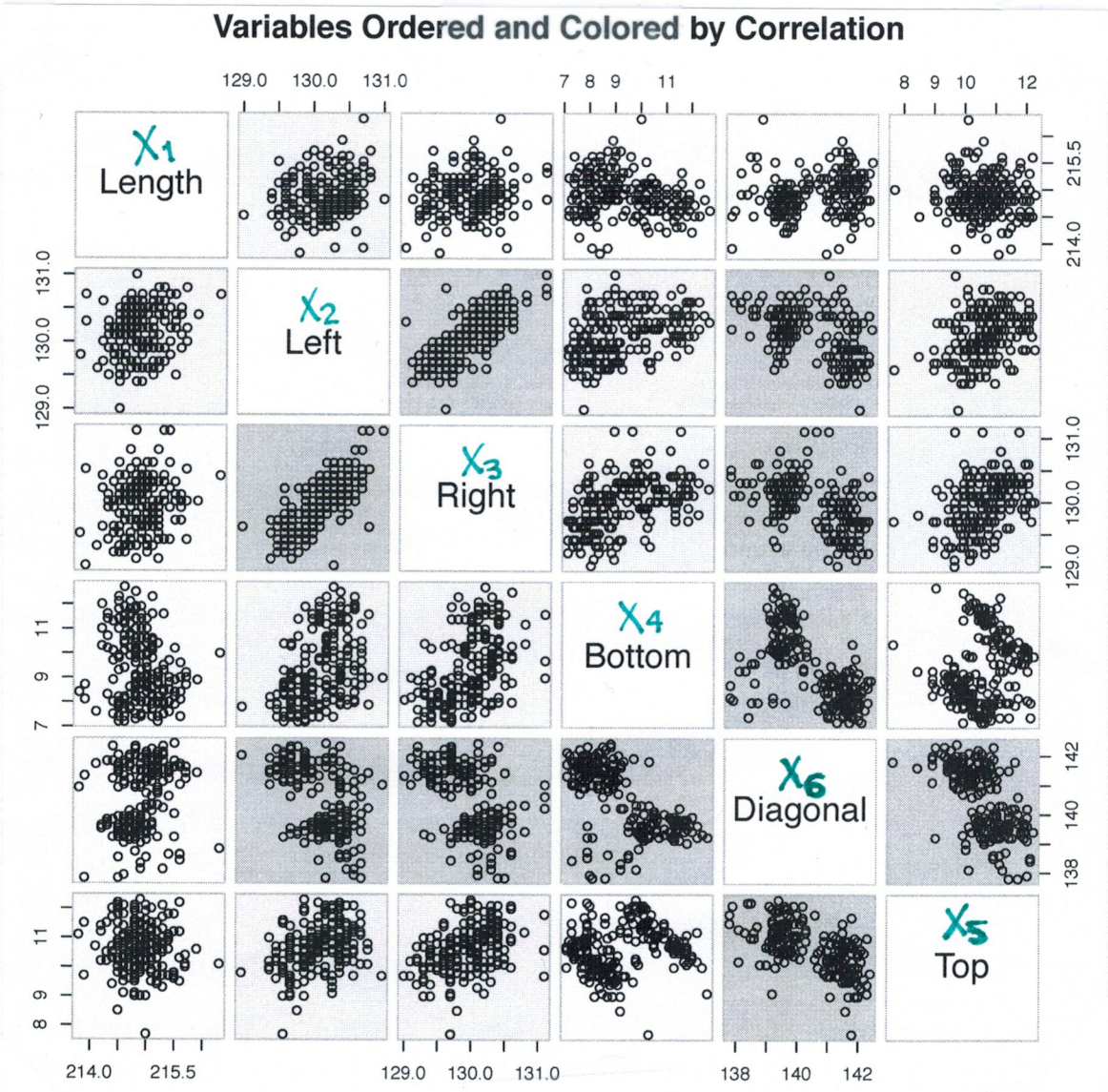


FIGURA D

```
notes1.R <- function(){
# produces a scatter plot of swiss banknote data
# first load package   uskewFactors v2.0, this contains
# Swiss banknote data.   Then invoque data(banknote)
# This function  requires package  gclus

pdf("scatter banknotes1.pdf")
library(uskewFactors)
data(banknote)
library(gclus)
dta <- banknote[c(1,2,3,4,5,6)]
dta.r <- abs(cor(dta))
dta.o <- order.single(dta.r)
dta.col <- dmat.color(dta.r)
cpairs(dta, dta.o, panel.colors=dta.col, gap=.5, main="Variables Ordered and Colored by
 Correlation" )
dev.off()

}
```

La figura D contiene los diagramas de dispersión de todos los vectores (X_i, X_j) ; $i \neq j$; $i, j \in \{1,2,\ldots,6\}$ para las componentes X_i y X_j de X. Observemos que el renglon 5 y la columna 5 de esta matríz muestran diagramas de dispersión para los cuales hay una clara separación de todos (los 200 individuos) los billetes en dos subgrupos. Este renglón y columna corresponden a la interacción de las componentes X_1, X_2, X_3, X_4 y X_5 con

La compontente X_6 = longitud de la diagonal del billete. Lo anterior nos hace pensar que es esta componente (esta característica de los billetes) la que nos puede ayudar a determinar 2 subpoblaciones[1]: Los billetes genoinos y los billetes falsos.

El problema de clasificación en la Estadística, consiste en determinar qué individuos en la muestra pertenecen a un subgrupo y cuáles a otro subgrupo. Además, es de interés estudiar si el método empleado servirá para clasificar nuevos individuos, provenientes de la misma población, que se observen a futuro. En principio, nosotros vamos a revisar metodología para clasificar en este curso.

(1) El nombre correcto aquí sería subgrupos porque al principio hablamos de una muestra

Al momento, nosotros intuímos que es la componente X_6 = longitud de la diagonal, la que nos permitirá clasificar[1], pero no tenemos (todavía) conocimiento de métodos de clasificación. En espera de aprender sobre métodos de clasificación posteriormente en el curso, supóngase que ya tenemos información de la muestra, sobre qué valores de $x_{1,6}, \ldots, x_{200,6}$ corresponden a billetes falsos y, que valores corresponden a billetes genuinos. Asumiendo que $x_{1,6}, \ldots, x_{100,6}$ corresponden a billetes verdaderos y $x_{101,6}, \ldots, x_{200,6}$ corresponden a billetes falsos, la función de R "MVAboxbank6.R" produce un gráfico de cajas para comparar estos dos subgrupos.

(1) Conocer esto, es producto de nuestras técnicas para gráficar, representar y explorar los datos. Este conocimiento se usará como insumo de los métodos de clasificación.