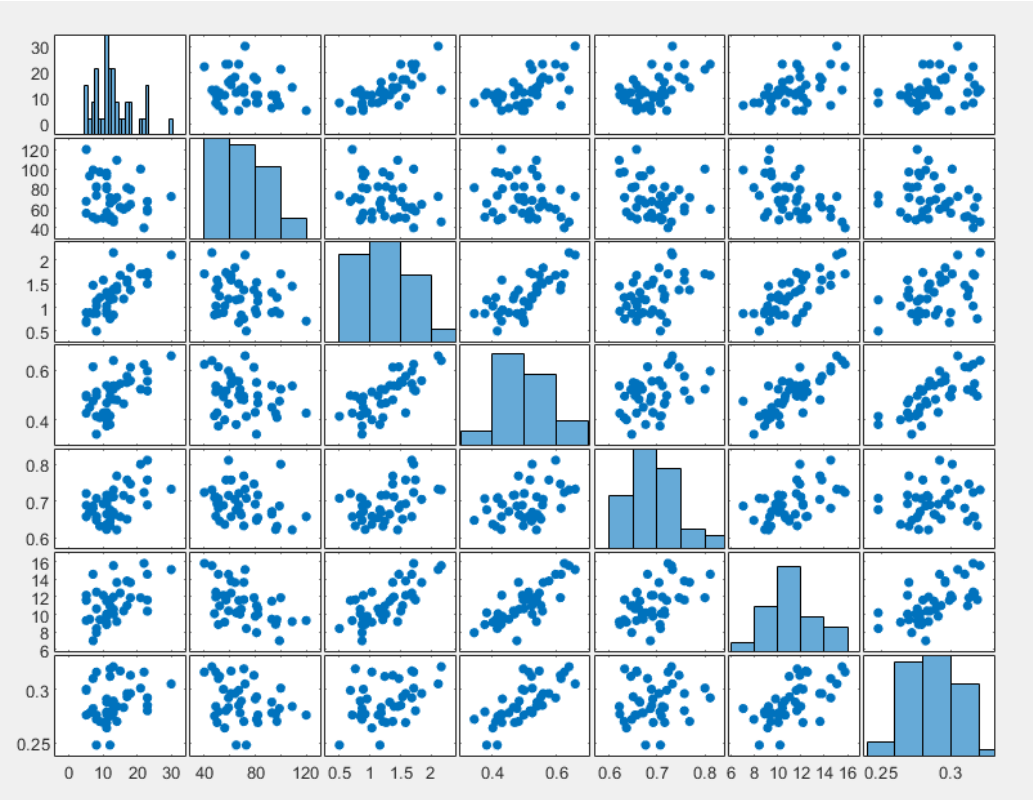
TERCERA TAREA

Daniel Aramburu y Guillermo Palomo G28

Clasificación jerárquica

Primeramente, vamos a recordar las variables a estudiar y a realizar las transformaciones pertinentes para conseguir hacer estas variables más normales. Nuestras variables son: “goles del máximo goleador del equipo” (X1), “número de tarjetas amarillas” (X2), “goles por partido” (X3), “% de posesión media” (X4), “% de salvadas del portero” (X5), “tiros del equipo por partido” (X6) y “% de presión exitoso” (X7).

Con Matlab observamos este gráfico de dispersión matricial:



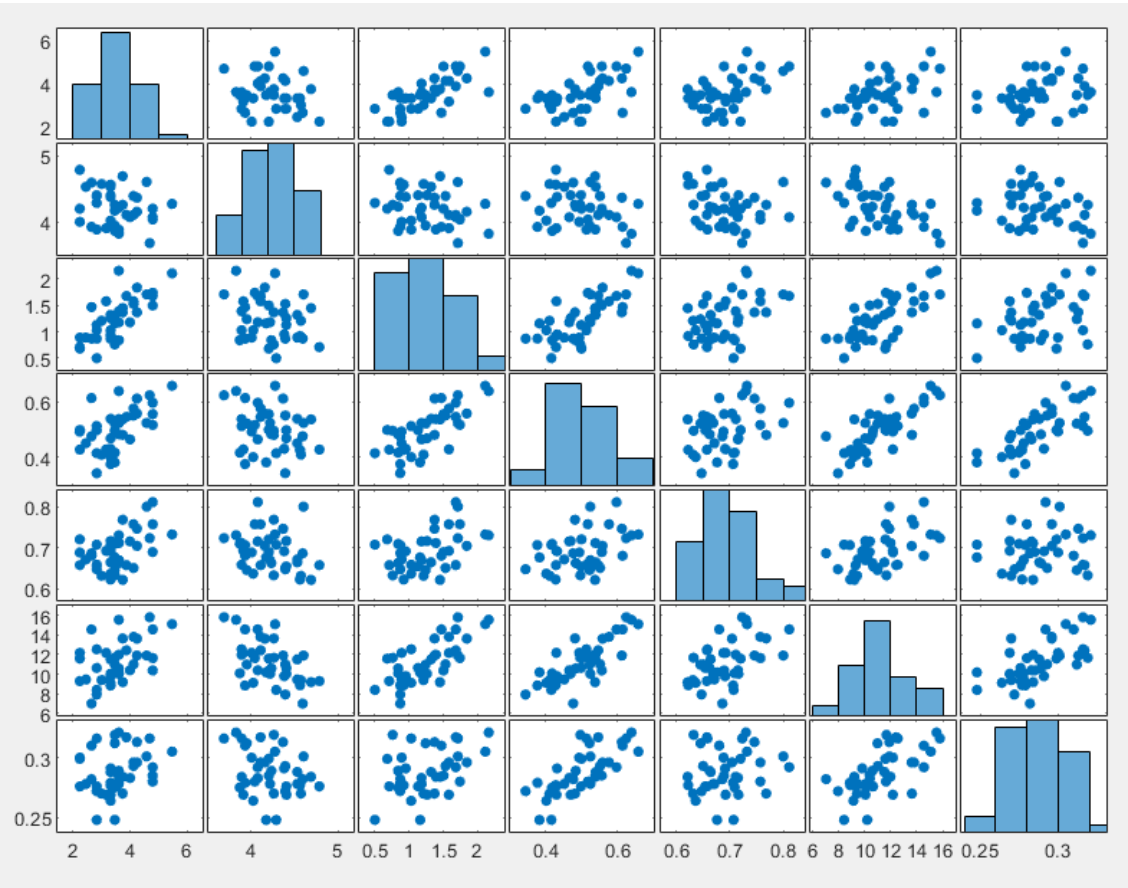
A priori, las variables que tienen más posibilidades de necesitar transformación son X1 y X2. Vamos a pasar a R para comprobar si se pueden considerar normales o no.

```
> shapiro.test(QMulti[,1])

        Shapiro-Wilk normality test

data:  QMulti[, 1]
W = 0.9321, p-value = 0.01887

> shapiro.test(QMulti[,2])

        Shapiro-Wilk normality test

data:  QMulti[, 2]
W = 0.95212, p-value = 0.08973

> shapiro.test(QMulti[,3])

        Shapiro-Wilk normality test

data:  QMulti[, 3]
W = 0.97301, p-value = 0.4458

> shapiro.test(QMulti[,4])

        Shapiro-Wilk normality test

data:  QMulti[, 4]
W = 0.98492, p-value = 0.8619
```

```
> shapiro.test(QMulti[,5])

        Shapiro-Wilk normality test

data:  QMulti[, 5]
W = 0.96095, p-value = 0.1805

> shapiro.test(QMulti[,6])

        Shapiro-Wilk normality test

data:  QMulti[, 6]
W = 0.97182, p-value = 0.4101

> shapiro.test(QMulti[,7])

        Shapiro-Wilk normality test

data:  QMulti[, 7]
W = 0.96502, p-value = 0.2477
```

Observamos por el p-valor que, efectivamente, debemos transformar X1 y X2 por lo que nuestras nuevas variables quedarán de esta manera:

```
X = [sqrt(X(:,1)) log(X(:,2)) X(:,3) X(:,4) X(:,5) X(:,6) X(:,7)];
```



Como vemos, ahora las variables se hallan más centradas. Podemos decir también que estas variables están corraladas como vemos en la siguiente matriz de correlaciones:

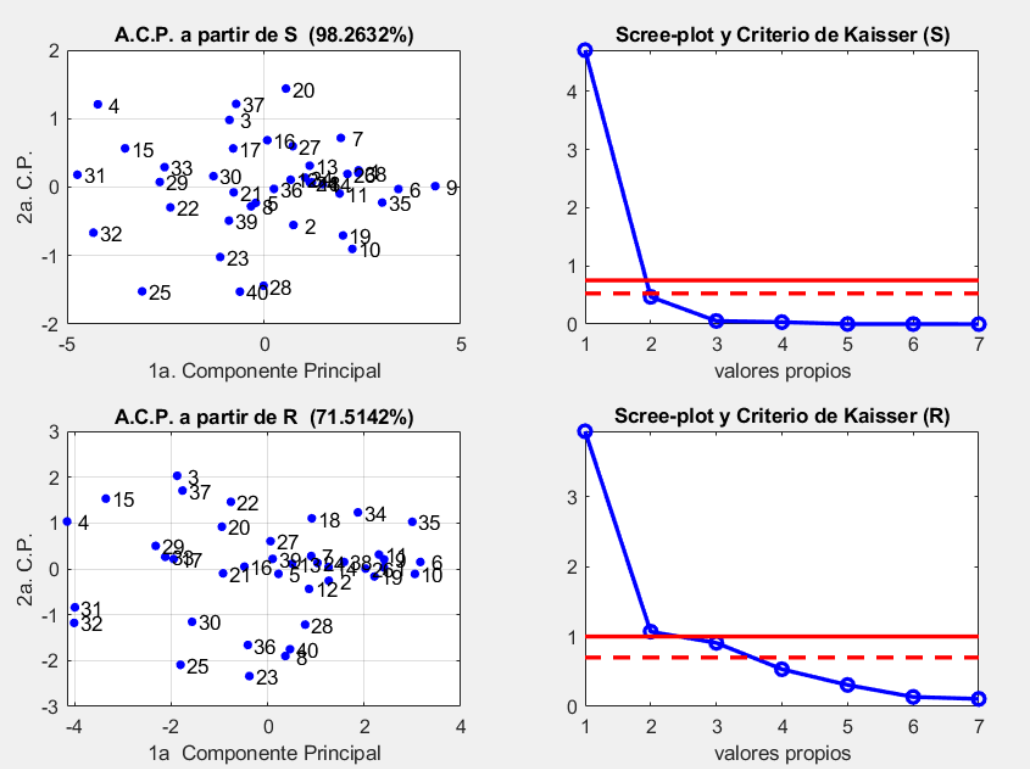
```
R =

    1.0000   -0.1912    0.7368    0.5599    0.4991    0.4799    0.2582
   -0.1912    1.0000   -0.2664   -0.2428   -0.2594   -0.5359   -0.3043
    0.7368   -0.2664    1.0000    0.7323    0.4383    0.7271    0.4111
    0.5599   -0.2428    0.7323    1.0000    0.3892    0.7896    0.7294
    0.4991   -0.2594    0.4383    0.3892    1.0000    0.5194    0.1238
    0.4799   -0.5359    0.7271    0.7896    0.5194    1.0000    0.6358
    0.2582   -0.3043    0.4111    0.7294    0.1238    0.6358    1.0000
```

Para poder utilizar el análisis de componentes principales vamos a observar las varianzas para saber si se necesita seguir con la matriz de correlaciones anterior o con la de covarianzas.

```
V =

    0.6065    0.0706    0.1596    0.0059    0.0022    4.5605    0.0003
```

Hay alguna diferencia notable entre varianzas por lo que procederemos a utilizar la matriz de correlaciones para calcular las componentes principales. Para llegar a una variabilidad explicada superior al 80% vamos a coger las 3 primeras componentes, coincidiendo con el criterio de la modificación de Jollife, con las que se consigue explicar el 84.51% de la variabilidad total como vemos abajo.

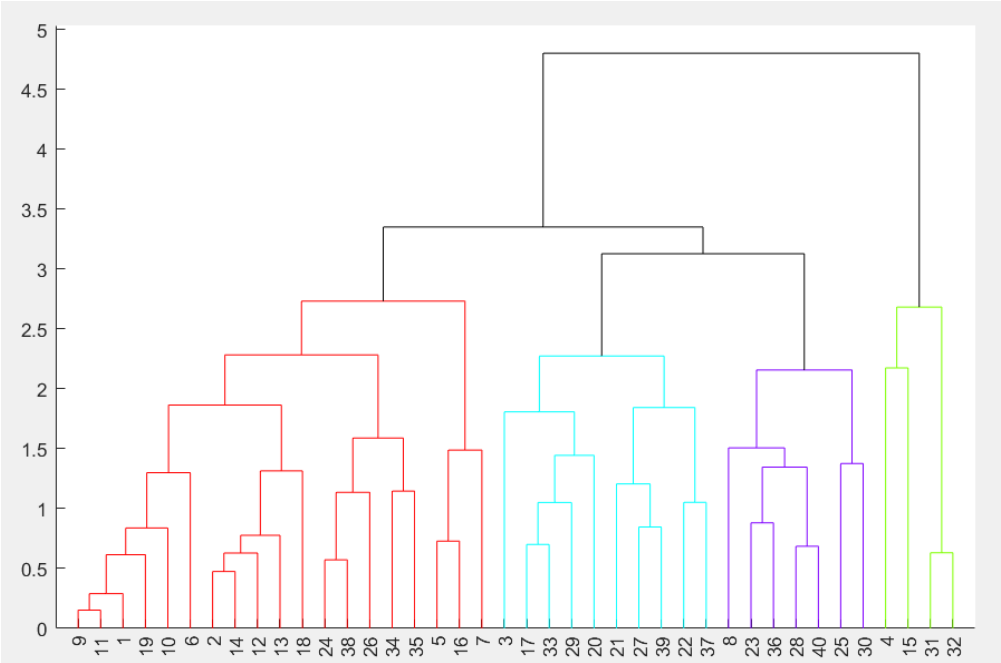
```
acum2 =

   56.2382   71.5142   84.5130   92.1153   96.5081   98.4584  100.0000
```



Calculamos ahora la distancia euclídea y probamos los 3 métodos de clasificación jerárquica eligiendo la que mayor correlación cofenética tenga.

c_max	0.6664
c_min	0.6274
c_UPGMA	0.7200

Por lo que vemos, el método UPGMA es el que menos distorsiona las distancias originales con un grado de coherencia del 72% por lo que lo utilizaremos para hacer el dendograma.

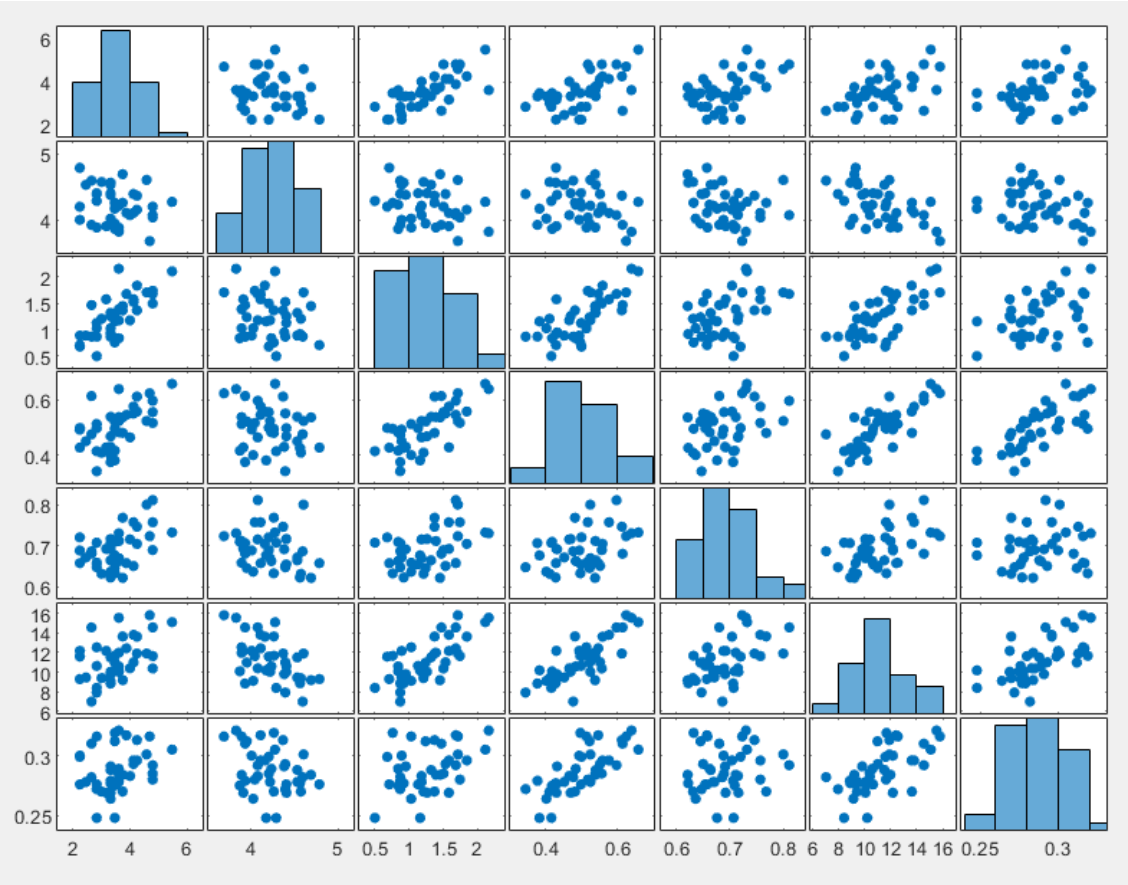


Al dendograma le hemos puesto un límite de distancia 3 para crear los grupos de colores, formando 4 grupos de colores coincidiendo en gran medida con nuestro análisis MDS de la práctica 2. El grupo verde lo componen los equipos con grandes estrellas y que mejor y más entretenido para el espectador juegan, por lo que suelen quedar arriba en la clasificación. El morado son los equipos sin grandes estrellas y que peor presionan por lo que suelen quedar de los últimos en la tabla de clasificación de sus respectivas ligas. El grupo azul lo conforman equipos un escalón por debajo del grupo verde en cuanto a sus características principales, quedando normalmente 4-8 en la clasificación y el grupo rojo lo conforman el resto de los equipos de las ligas que no destacan en nada muy significativo.

Clasificación no jerárquica

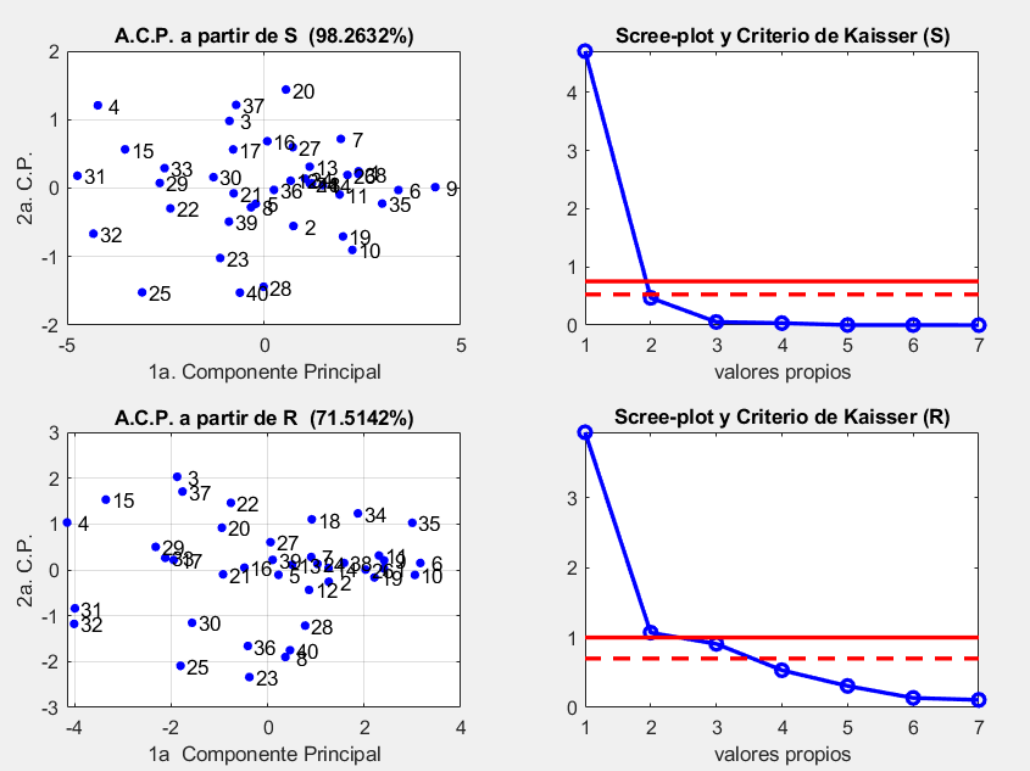
En cuanto a la clasificación no jerárquica estudiaremos las mismas variables ya que queremos observar y analizar las diferencias y similitudes con respecto a la jerárquica. Para realizar esta clasificación, primero hemos realizado transformaciones no lineales en dos variables y hemos dejado las demás iguales. Esto lo realizamos con el objetivo de que al estar las variables finales formadas por variables transformadas centramos todo lo que podemos los datos. Las transformaciones mencionadas son las siguientes:

```
X = [sqrt(X(:,1)) log(X(:,2)) X(:,3) X(:,4) X(:,5) X(:,6) X(:,7)];
```



Al observar el gráfico nos damos cuenta de que no debemos aplicar el algoritmo de k-medias sobre los datos de manera directa mediante la distancia euclídea, ya que existen fuertes correlaciones entre algunas variables. Es por ello por lo que realizaremos el cálculo de los componentes principales para posteriormente realizar el algoritmo de k-medias sobre ellos.

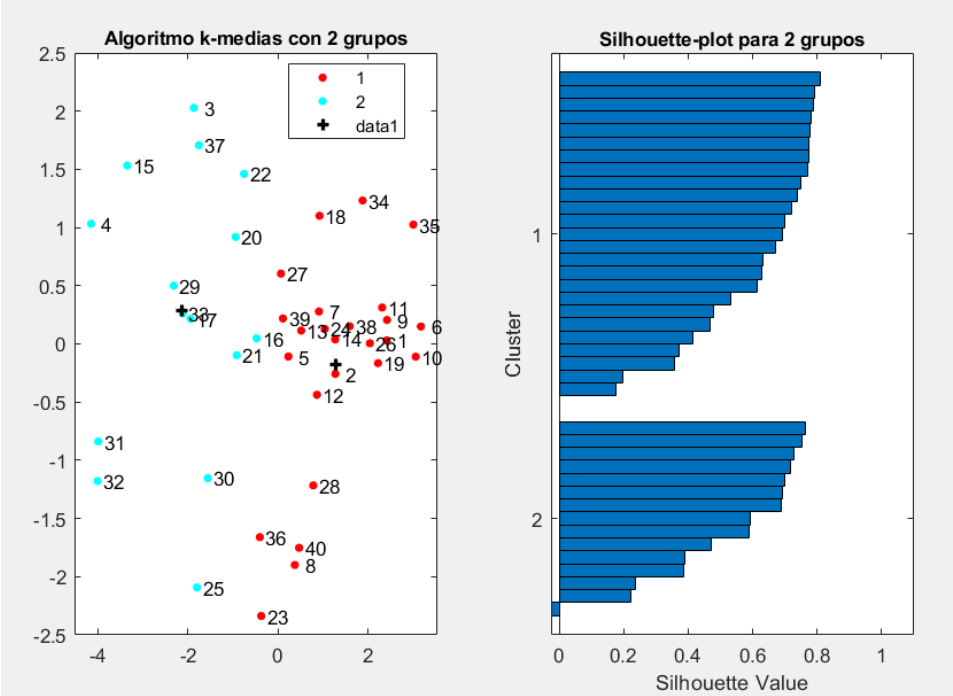
Debido a la diferencia entre varianzas, usamos la matriz de correlaciones para el cálculo de las componentes principales, al igual que hicimos anteriormente en la clasificación jerárquica. Dentro de las componentes vamos a escoger las tres primeras que explican un 84,51% de la variabilidad.



Este método se encarga de agrupar los individuos según la variabilidad, generando conglomerados externamente heterogéneos e internamente homogéneos.

A continuación, empezamos con el algoritmo de k-medias, ya que tenemos las componentes principales decididas. Dicho algoritmo trata de agrupar individuos según la variabilidad entre grupos generando de esta manera conglomerados.

Vamos a calcular los puntos medios(centroides) de cada conglomerado para posteriormente calcular la distancia euclídea de cada elemento a los distintos centroides.




```
C =

    1.2798   -0.1739   -0.0555
   -2.1330    0.2898    0.0925

s =

    0.5829
```

Respecto al primer conglomerado, las coordenadas de los centroides son (1.2798, -2.1330), y respecto al segundo, las coordenadas son (-0.1739,0.2898). Además, observamos que la silueta de la media es de 0.5829 en nuestro caso.

En el gráfico de la derecha podemos observar dos grupos, el segundo de ellos con menos individuos que el primero. Por último, hay que decir que en el segundo grupo existe una silueta negativa, lo que quiere decir que se debería agrupar con el conglomerado de al lado.

Una vez terminado el proceso anterior, procedemos a calcular las medias y medianas:

```
Medias =

    3.1195    4.2769    1.0144    0.4570    0.6710   10.0684    0.2817
    4.2227    4.1304    1.6560    0.5727    0.7277   13.1800    0.2986
```

La única variable que difiere de las demás es la X2 “número de tarjetas amarillas”, que es una de las variables transformadas. Todas las demás son mayores en el grupo 2 que en el grupo 1.

```
Medianas =

    3.3166    4.2627    0.9500    0.4650    0.6690   10.0300    0.2780
    4.2426    4.1109    1.6800    0.5580    0.7300   13.6100    0.2960
```

La distribución de las medianas es muy similar a la de las medias, por lo que creemos que el grupo 1 está formado por los equipos que juegan peor y consiguen más tarjetas amarillas; y el grupo 2 por los equipos que juegan mejor y por lo tanto no hacen tantas faltas que conlleven tarjetas amarillas.

Conclusión

En conclusión, estamos satisfechos con los resultados, ya que son coherentes debido a que obtenemos las mismas conclusiones a partir de dos clasificaciones diferentes, como son la jerárquica y la no jerárquica. Por un lado, en la jerárquica se forman 4 grupos mayormente basados en un estilo de juego limpio y entretenido, y como podemos comprobar, en la no jerárquica sucede lo mismo formando 2 grupos más amplios.

Anexo

```
X = MultiLiga{:,7:13};

plotmatrix(X);

X = [sqrt(X(:,1)) log(X(:,2)) X(:,3) X(:,4) X(:,5) X(:,6) X(:,7)];

plotmatrix(X);

R = corr(X);
V = var(X);

[T1,Y1,acum1,T2,Y2,acum2]=comp2(X);

Y2 = Y2 (:,1:3);

Y = pdist(Y2,'euclidean');

Z_min = linkage(Y,'single');

Z_max = linkage(Y,'complete');

Z_UPGMA = linkage(Y,'average');


c_UPGMA = cophenet(Z_UPGMA,Y);

c_min = cophenet(Z_min,Y);
```

```
c_max = cophenet(Z_max,Y);


dendrogram(Z_UPGMA,0,'colorthreshold',3);



[C,s,IDX]=kmedias2(Y2,6);

C;
s;

Medias = splitapply(@mean,X,IDX);

Medianas = splitapply(@median,X,IDX);
```