SEGUNDA TAREA

Daniel Aramburu y Guillermo Palomo G28

Recordatorio tarea 1

Nuestra base de datos contiene información sobre distintas variables futbolísticas de La Liga y la Premier League de la temporada 20/21. Las variables cuantitativas son “goles del máximo goleador del equipo”, “número de tarjetas amarillas”, “goles por partido”, “% de posesión media”, “% de salvadas del portero”, “tiros del equipo por partido” y “% de presión exitoso”. Podemos observar cómo las unidades de varias de estas variables son diferentes entre sí, aunque luego profundizaremos más. Las variables binarias son “liga” y “recién ascendido” y las categóricas multi-estado son “puestos importantes” y “rango de edad media del equipo”.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Equipo	Puesto	Rango de edad media	Puestos importantes	Liga	Recien ascendido	Goles del max goleador	Tarjetas amarillas	Goles por partido	Posesion media	% de salvadas	Tiros por partido	% de presion exitoso
2	Alaves	16	28,1-29	Media tabla	0	0	11	96	0,92	0,427	0,622	8,87	0,277
3	Athletic Club	10	26,1-27	Media tabla	0	0	8	82	1,13	0,498	0,654	10,63	0,276
4	Atletico Madrid	1	27,1-28	Champions	0	0	21	100	1,71	0,524	0,800	11,89	0,301
5	Barcelona	3	26,1-27	Champions	0	0	30	72	2,11	0,658	0,732	15,08	0,305
6	Betis	6	27,1-28	Europa League	0	0	11	93	1,32	0,539	0,664	11,50	0,288
7	Cadiz	12	28,1-29	Media tabla	0	1	8	81	0,87	0,343	0,647	7,92	0,272
8	Celta Vigo	8	26,1-27	Media tabla	0	0	14	109	1,45	0,537	0,621	9,18	0,284
9	Eibar	20	28,1-29	Descenso	0	0	12	71	0,76	0,496	0,632	11,66	0,319
10	Elche	17	28,1-29	Media tabla	0	1	7	99	0,87	0,476	0,686	7,00	0,282
11	Getafe	15	26,1-27	Media tabla	0	0	5	120	0,71	0,428	0,657	9,29	0,276
12	Granada	9	27,1-28	Media tabla	0	0	9	97	1,21	0,410	0,631	9,39	0,270
13	Huesca	18	27,1-28	Descenso	0	1	13	69	0,84	0,483	0,662	10,58	0,289
14	Levante	14	27,1-28	Media tabla	0	0	13	68	1,18	0,524	0,669	10,03	0,278
15	Osasuna	11	27,1-28	Media tabla	0	0	11	80	0,95	0,431	0,691	9,79	0,291
16	Real Madrid	2	27,1-28	Champions	0	0	23	59	1,68	0,597	0,811	14,55	0,292
17	Real Sociedad	5	25-26	Europa League	0	0	17	81	1,53	0,551	0,650	11,00	0,293
18	Sevilla	4	28,1-29	Champions	0	0	18	79	1,37	0,612	0,746	11,89	0,313
19	Valencia	13	25-26	Media tabla	0	0	11	82	1,26	0,469	0,716	10,03	0,269
20	Valladolid	19	28,1-29	Descenso	0	0	6	93	0,89	0,451	0,671	9,45	0,278
21	Villarreal	7	27,1-28	Media tabla	0	0	23	67	1,50	0,556	0,689	10,37	0,285
22	Arsenal	8	25-26	Media tabla	1	0	13	49	1,39	0,538	0,711	11,97	0,284
23	Aston Villa	11	25-26	Media tabla	1	0	14	71	1,37	0,481	0,768	13,63	0,270
24	Brighton	16	25-26	Media tabla	1	0	8	49	1,03	0,513	0,658	12,53	0,316
25	Burnley	17	28,1-29	Media tabla	1	0	12	48	0,84	0,417	0,709	10,08	0,275
26	Chelsea	4	25-26	Champions	1	0	7	51	1,47	0,614	0,680	14,55	0,310
27	Crystal Palace	14	29,1-30	Media tabla	1	0	11	56	1,03	0,401	0,636	9,11	0,264
28	Everton	10	26,1-27	Media tabla	1	0	16	59	1,18	0,465	0,715	10,39	0,283
29	Fulham	18	25-26	Descenso	1	1	5	67	0,68	0,499	0,720	11,58	0,299
30	Leeds United	9	26,1-27	Media tabla	1	1	17	61	1,58	0,576	0,757	13,79	0,296
31	Leicester City	5	26,1-27	Europa League	1	0	15	61	1,68	0,546	0,657	12,42	0,317
32	Liverpool	3	26,1-27	Champions	1	0	22	40	1,71	0,624	0,723	15,79	0,316
33	Manchester City	1	26,1-27	Champions	1	0	13	46	2,16	0,639	0,730	15,53	0,321
34	Manchester Utd	2	25-26	Champions	1	0	18	64	1,84	0,558	0,704	13,61	0,296
35	Newcastle Utd	12	27,1-28	Media tabla	1	0	12	65	1,16	0,382	0,676	10,18	0,248
36	Sheffield Utd	20	26,1-27	Descenso	1	0	8	73	0,50	0,415	0,707	8,39	0,248
37	Southampton	15	26,1-27	Media tabla	1	0	12	52	1,24	0,522	0,645	10,97	0,312
38	Tottenham	7	27,1-28	Media tabla	1	0	23	57	1,74	0,517	0,757	11,63	0,280
39	West Brom	19	26,1-27	Descenso	1	1	11	51	0,87	0,376	0,706	8,84	0,279
40	West Ham	6	27,1-28	Europa League	1	0	10	50	1,58	0,429	0,692	12,16	0,269
41	Wolves	13	26,1-27	Media tabla	1	0	5	55	0,89	0,493	0,688	12,16	0,300

Vamos a realizar los análisis de componentes y de coordenadas principales (ACP y MDS).

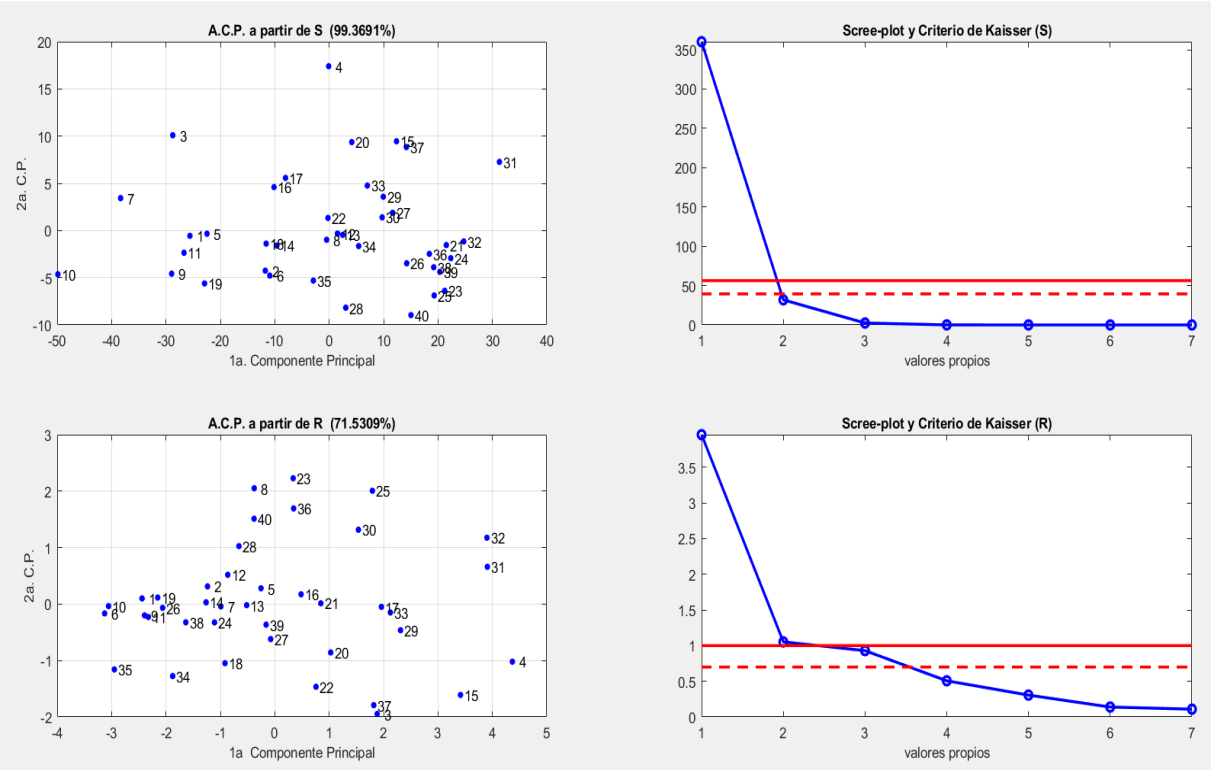
Análisis de Componentes Principales ACP

Para realizar este análisis se utilizan las variables cuantitativas mencionadas anteriormente. Podemos ver que las unidades de algunas variables difieren entre sí, yendo de porcentajes, a tiros por partido, goles promedio y hasta tarjetas amarillas en la temporada. Esto puede suponer un problema para las varianzas haciendo que las más pequeñas contribuyan muy poco en los primeros ejes principales. También comprobamos la eta^2 para ver si hay relaciones lineales entre las variables.

```
VARS =

   33.2506  366.7635    0.1596    0.0059    0.0022    4.5605    0.0003

>> eta2Z

eta2Z =

    0.9908
```

Las variables en orden 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). Como cabía esperar las varianzas de las variables X1, X2 y X6 son significativamente superiores a las demás por lo que el ACP se llevará a cabo a partir de la matriz de correlaciones. Para esto se estandarizan las variables a media cero y varianza uno, pasando a utilizar la matriz de covarianzas de las variables estandarizadas.

La eta^2 de 0.9908 tan cercano a 1 indica que las variables están muy relacionadas por lo que el análisis se podría realizar con pocas componentes principales.




```
acum2 =

   56.4944   71.5309   84.8118   92.0550   96.4414   98.4351  100.0000
```

Los dos gráficos de la fila de arriba nos señalan como sería el análisis utilizando la matriz de covarianzas en vez de la de correlaciones. En este, solo la primera variable ya explica el 99.369% de la variabilidad por lo que este análisis no nos es muy útil.

Los dos gráficos de la fila de abajo están basados en la matriz de correlaciones. Usando el criterio de Kaisser vemos como las dos primeras componentes principales explican el 71.53% de la variabilidad, y como vemos en el screeplot con la línea discontinua, según la modificación de Jollife, incluiríamos las 3 primeras variables consiguiendo un 84.881% de variabilidad total explicada.

Para continuar el análisis utilizaremos el criterio de Kaisser con las 2 componentes principales.

```
>> T2

T2 =

    0.3748   -0.4516    0.1746    0.3932    0.6011   -0.0741    0.3220
   -0.2394   -0.2441    0.8281   -0.3420   -0.1548    0.0977    0.2164
    0.4300   -0.1912    0.1972    0.4021   -0.4642    0.4536   -0.3959
    0.4463    0.1776    0.2973   -0.1122   -0.0694   -0.7421   -0.3358
    0.3059   -0.5317   -0.3018   -0.6876    0.0692    0.1235   -0.1994
    0.4579    0.1594   -0.1412   -0.1171   -0.4420   -0.0152    0.7319
    0.3390    0.6002    0.2132   -0.2601    0.4407    0.4615   -0.0676
```

<u>La ecuación de la primera componente principal es:</u>

Y1 = 0.3748 * GolesMaxGoleador - 0.2394 * NumTjtasAmarillas + 0.43 * GolesxPartido + 0.4463 * %PosesionMedia + 0.3059 * %SalvadasPortero + 0.4579 * TirosxPartido + 0.339 * %PresionExitoso

La contribución de las variables originales (en valor absoluto) no suele diferir mucho, estando normalmente entre 0.3-0.45. La variable de tarjetas amarillas contribuye de forma negativa mientras que todas las demás lo hacen de forma positiva, siendo las más influyentes las de tiros por partido y porcentaje de posesión media.

Aquellos equipos cuyos máximos goleadores marquen muchos goles, tengan pocas tarjetas amarillas, metan muchos goles por partido, tengan altos porcentajes de posesión del balón, su portero haga muchas paradas, tiren muchos tiros por partido y tengan altos porcentajes de presión exitosa tendrán una primera componente principal con un valor elevado.

A su vez, para aquellos equipos cuyos máximos goleadores marquen pocos goles, tengan muchas tarjetas amarillas, metan pocos goles por partido, tengan bajos porcentajes de posesión del balón, su portero haga pocas paradas, tiren pocos tiros por partido y tengan bajos porcentajes de presión exitosa tendrán una primera componente principal con un valor bajo.

Por todo lo anterior, esta componente nos ordena los equipos de los que peor juegan a los que mejor juegan. Podría usarse como indicador de nivel del juego del equipo teniendo en cuenta las individualidades como lo bueno que sea el máximo goleador o el portero, y lo sucio que juegue el equipo por el número de tarjetas amarillas. Este indicador es interesante desde el punto de vista del espectador ya que los mejores equipos según esta componente serán los más entretenidos de ver. ¿Serán también los que mejor queden en la clasificación?

Por ejemplo, los equipos con mejor juego (derecha del gráfico de individuos) son el FC Barcelona(4), el Manchester City(32) y el Liverpool(31), que quedaron 3ro, 1ro y 3ro en sus respectivas ligas. Los equipos con peor juego, y entonces, más aburridos de ver (izquierda en el gráfico de individuos), son el Cádiz(6), el Getafe(10) y el Sheffield United(35), que quedaron 12avo, 15avo y 20avo en sus respectivas ligas. Esto nos hace ver que, un equipo que juegue bien y tenga estrellas, es muy posible que quede entre los 3 primeros de la clasificación, mientras que un equipo sin estrellas y que juegue sucio, en La Liga, no quedará de los últimos, pero en la Premier League es más común que acabe descendiendo.

La ecuación de la segunda componente principal es:

Y2 = - 0.4516 * GolesMaxGoleador - 0.2441 * NumTjtasAmarillas - 0.1912 * GolesxPartido + 0.1776 * %PosesionMedia - 0.5317 * %SalvadasPortero + 0.1594 * TirosxPartido + 0.6002 * %PresionExitoso

En nuestra segunda componente principal, hay 4 variables originales contribuyendo negativamente y 3 positivamente. Las más influyentes (en valor absoluto) son las de porcentajes de presión exitosos, salvadas de porteros y goles del máximo goleador, siendo la primera positiva y la segunda y tercera negativas.

Aquellos equipos con goleadores a los que les cuesta marcar, cuyo portero no pare los tiros y presionen bien, tendrán un valor elevado en esta componente. Al contrario, equipos con goleadores y porteros buenos y que presionen mal, tendrán un valor bajo.

Por ello, esta componente ordena a los equipos principalmente según el riesgo que toman al presionar alto al rival y secundariamente según el nivel de las estrellas del equipo. Abajo veremos los equipos que peor presionan y tienen estrellas de alto rendimiento y arriba los equipos que mejor presionan, pero sus estrellas son de menor nivel futbolístico. Podemos decir que es un indicador de la organización del equipo en la presión alta y la implicación de las estrellas del equipo en esta.

Los individuos más destacados según la segunda componente (arriba en el gráfico de individuos) son, el Brighton(23), el Eibar(8) y el Chelsea(25), que quedaron 16avo, 20avo y 4to en sus respectivas ligas. Estos equipos coinciden en tener porteros y goleadores mediocres y buena presión adelantada. Lo que diferencia a los 2 primeros del Chelsea, que quedó muy arriba en la clasificación, puede ser que fuera que los goleadores en el equipo estuvieran muy repartidos y que su defensa en campo propio fuera mucho mejor que la de los otros equipos, pero se necesitarían más datos para afirmar esto.

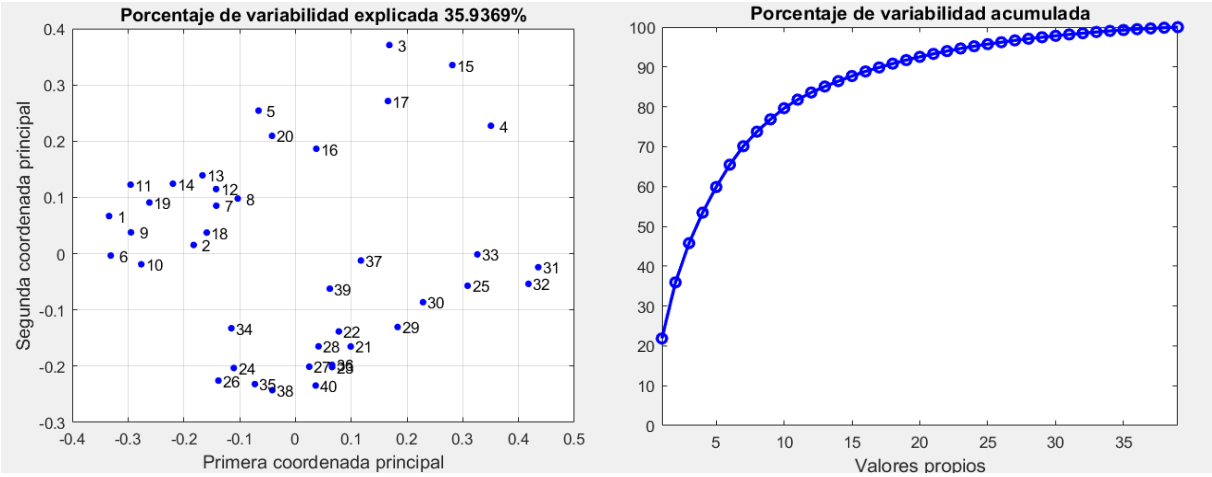
Por otro lado, los individuos más destacados por una mala presión alta pero con grandes estrellas (abajo en el gráfico de individuos), son el Atlético de Madrid(3), el Tottenham(37) y

el Real Madrid(15), que quedaron 1ro, 7mo y 2do en sus respectivas ligas. Estos tienen los mejores delanteros y porteros de las competiciones y por ello no les hace falta presionar tan bien para acabar en los puestos altos de la clasificación.

Análisis de Coordenadas Principales MDS

Para este análisis hemos cambiado de orden las variables de la matriz utilizada en el ACP y hemos añadido las binarias y categóricas multi-estado. El nuevo orden para esta matriz es, primero las 7 variables cuantitativas, “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); las 2 siguientes son las binarias de “liga”(X8) y “recién ascendido”(X9); y finalmente las categóricas multi-estado son 2, “puestos importantes”(X10) y “rango de edad media”(X11).

Con la matriz de cuadrados de distancias sacamos la matriz de coordenadas principales de la cual obtenemos estos gráficos.

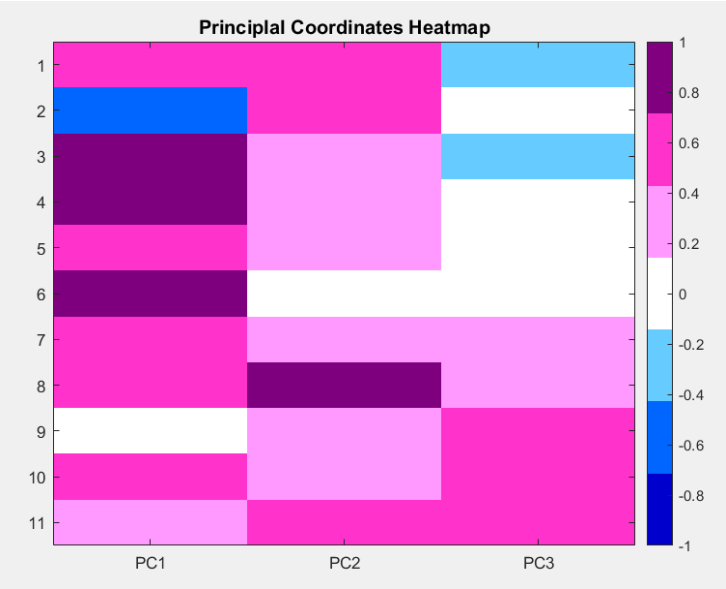


Las dos primeras coordenadas principales representan el 35.9369% de la variabilidad explicada, mientras que para llegar a un 80% necesitaríamos 11 coordenadas como vemos en el segundo gráfico y en esta variable:

```
>> acum(1:11)

ans =

   21.8506   35.9369   45.7745   53.4465   59.8672   65.4719   70.1102   73.7694   76.8721   79.6516   81.8045
```

A continuación, para interpretar los ejes vamos a calcular, para las variables cuantitativas los coeficientes de correlación de Pearson y para las nominales la V de Cramer.

```
correlaciones =

    0.5835    0.4648   -0.2288
   -0.5980    0.5195    0.1198
    0.7596    0.3430   -0.3183
    0.7875    0.3742    0.0343
    0.5920    0.1510   -0.0722
    0.9081    0.0905   -0.0490
    0.6240    0.1806    0.3115
    0.6481    0.9055    0.3162
    0.1400    0.3131    0.5774
    0.5654    0.3641    0.4939
    0.3815    0.4726    0.4886
```

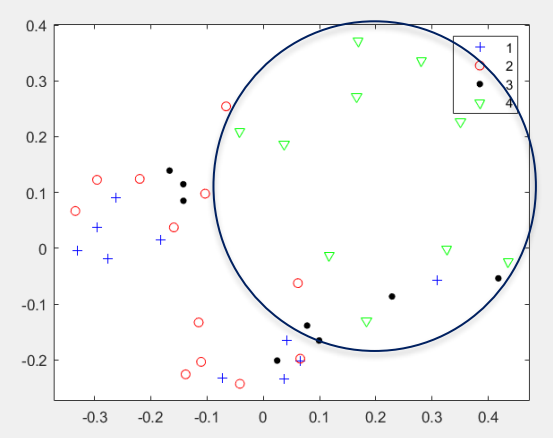


Las variables más influyentes en el primer eje principal son “goles por partido”(X3), “% de posesión media”(X4) y “tiros del equipo por partido”(X6) con correlación positiva y “número de tarjetas amarillas”(X2) con correlación negativa. Encontraremos a la derecha del primer eje los valores altos de las variables con correlación positiva y con niveles bajos de la variable con correlación negativa.

En el segundo eje principal la única variable influyente es la de “liga”(X8), con correlación positiva. Veremos entonces a los equipos de La Liga en la zona alta del segundo eje.

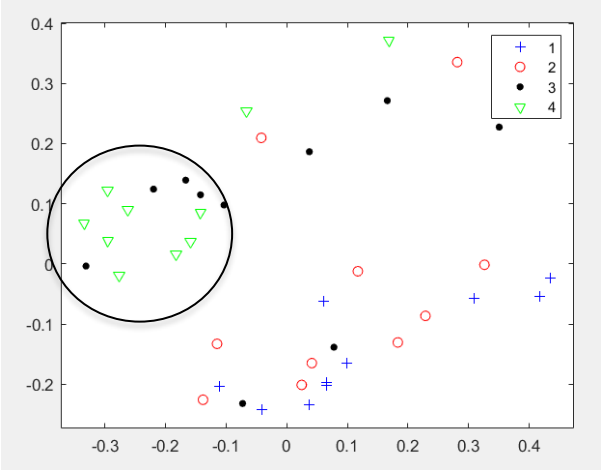
Para finalizar vamos a realizar el análisis de identificación de grupos y descripción de perfiles.

Goles del máximo goleador del equipo (X1)



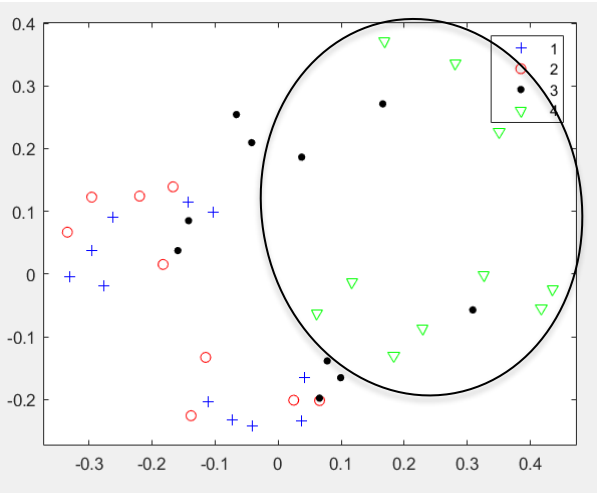
Los equipos cuyos máximos goleadores meten más goles se encuentran en la parte superior derecha. Nos ayuda a construir el perfil de los equipos que se acaban metiendo en puestos de Champions, es decir, los primeros 4 de cada liga.

Número de tarjetas amarillas (X2)

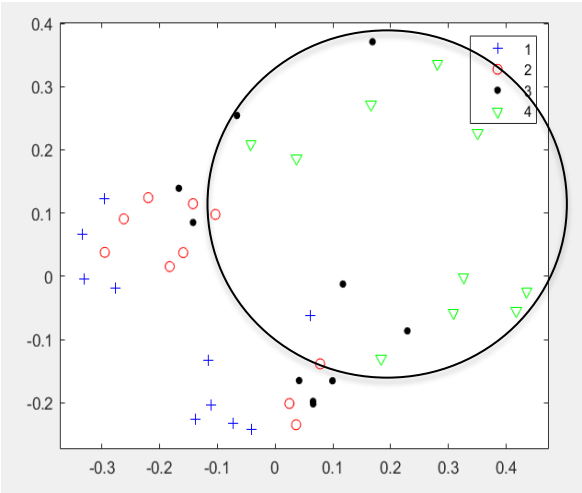


Aquí podemos ver como los equipos con más tarjetas amarillas suelen quedar peor en la clasificación y son de La Liga.

Goles por partido (X3)



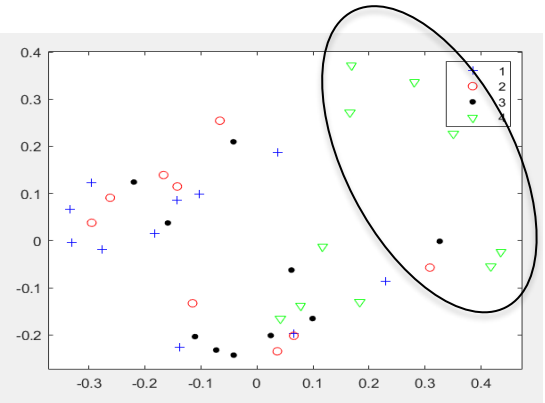
% de posesión media (X4)



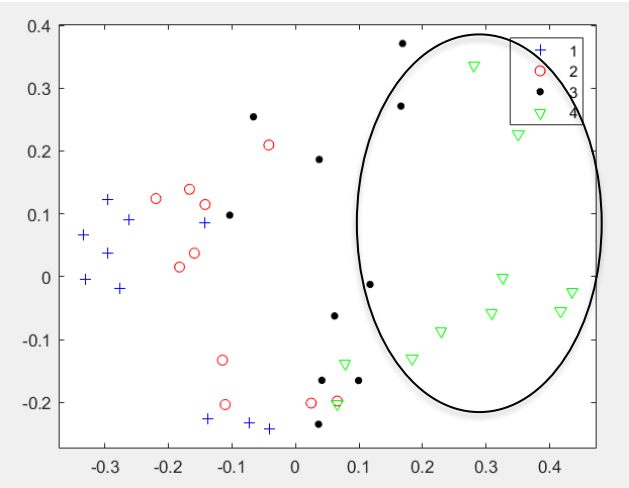
Seguimos observando como en el sector derecho, donde quedan los puestos de Champions, se encuentran los equipos con más goles por partido y más posesión independientemente de la liga que sea.

% de salvadas del portero (X5)

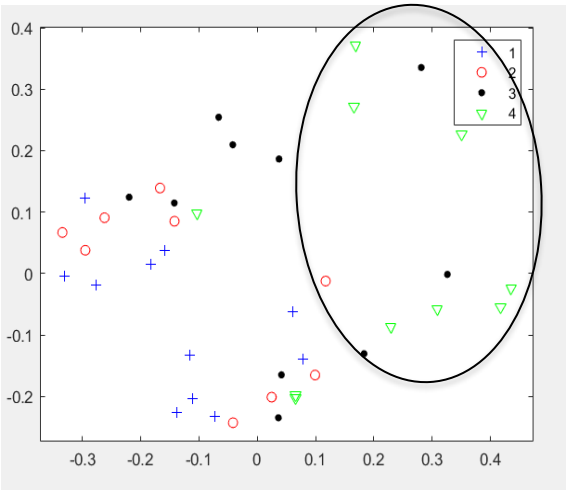
Con las salvadas de los porteros la tendencia anterior continúa, aunque no tan claramente como en variables anteriores como cabía esperar según el mapa de calor.



Tiros del equipo por partido (X6)

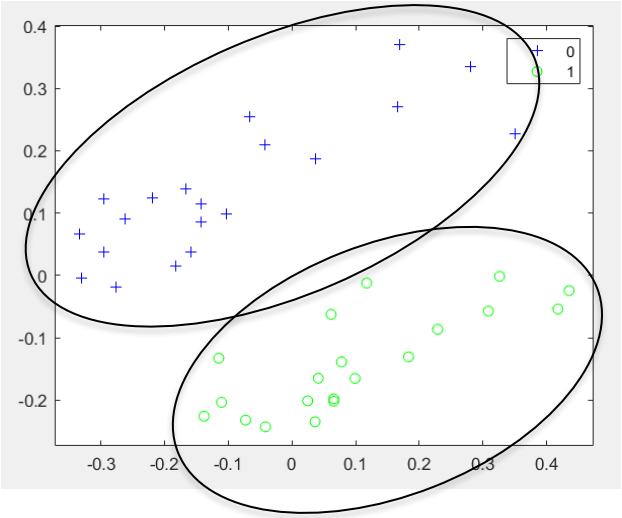


% de presión exitoso (X7)

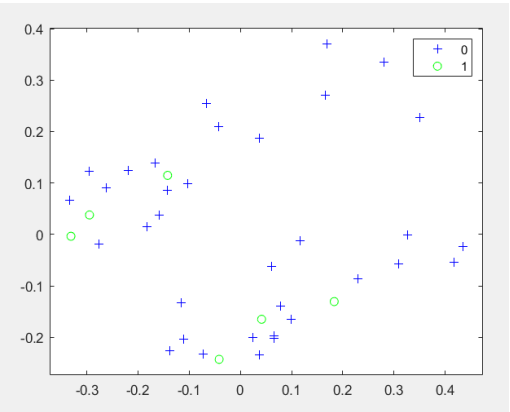


Seguimos viendo como a la derecha se encuentran los que más tiros a puerta hacen y los que mejor presionan.

Liga (X8)



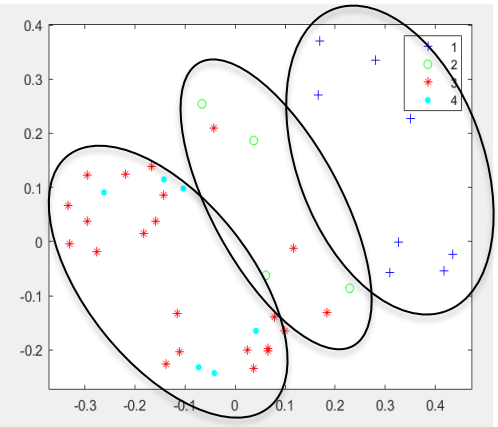
Aquí ya vemos dos grupos claramente diferenciados. En azul y en la parte alta, La Liga y abajo y en verde, la Premier League.



Recién ascendido (X9)

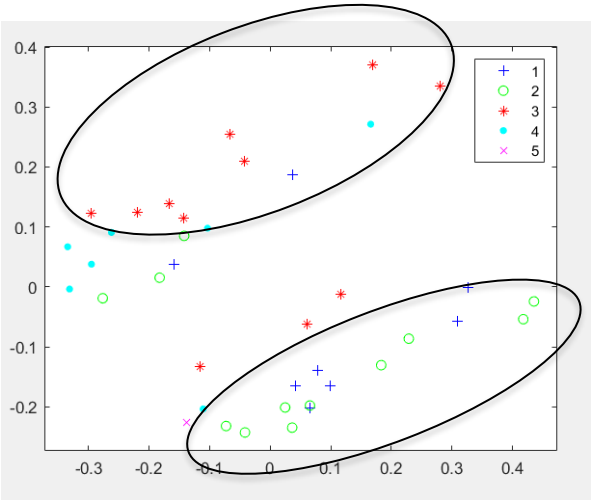
En este caso la variable no es muy significativa y no se ven grupos claros.

Puestos importantes (X10)



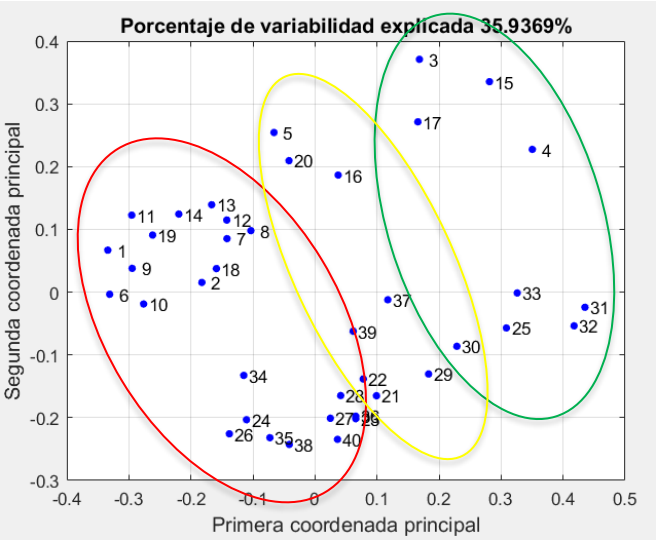
Se diferencian claramente los equipos en puestos de Champions (cruces azules) de los demás equipos, dejando ver que existe una gran superioridad entre ellos y el resto de los competidores. Aun así, como ya analizaremos después, podríamos distinguir 2 grupos más a parte de los equipos en puestos de Champions, siendo estos los de Europa League (verde) y finalmente los equipos de media tabla y colistas (rojo y azul claro).

Rango de edad media (X11)



De este gráfico podemos decir, aunque no muy claramente, que los equipos de La Liga suelen estar normalmente formados por jugadores más veteranos que los de la Premier League, estando en rojo arriba los equipos de La Liga de 27-28 años y abajo en verde los de la Premier League con 26-27 años de media.

Conclusión:



Creemos que los 3 perfiles más diferenciados y significativos son los señalados, aunque se podrían haber tenido en cuenta también otras variables como la liga o la edad para hacer otros perfiles diferentes.

Perfil 1 (verde): Este grupo lo forman los equipos que mejor y más entretenido para el espectador juegan, es decir, tiran más, meten más goles, presionan mejor etc. Son los equipos suelen acabar siempre entre los 4 primeros sin distinción de liga.

Perfil 2 (amarillo): Este grupo es el menos numeroso y lo forman los equipos que están un poco por debajo en nivel de juego comparado con los primeros, pero juegan limpio y consiguen acabar en la parte media-alta de la tabla de clasificación de las ligas.

Perfil 3 (rojo): Estos son los equipos que pelean por la permanencia en las ligas. Suelen ser los que juegan más sucio haciendo más amarillas, aunque esto se da más en equipos de La Liga solamente. En general estos equipos no tienen estrellas que marquen la diferencia y tienen que recurrir a estilos de fútbol más defensivos y no muy técnicos.

Apéndice de código

ACP

```
#Variables numéricas
Z = MultiLiga{:,7:13};

VARS = var(Z);

RZ = corr(Z);
eta2Z = 1 - det(RZ);

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

MDS

```
W = MultiLigaMDS{:,3:13};

[n,p] = size(W);

p1 = 7;
p2 = 2;
p3 = 2;

S = gower2(W,p1,p2,p3);
D2 = ones(size(S)) - S;

[Y,vaps,percent,acum] = coorp(D2);

pcuant = p1;
pnominal = p2 + p3;
correlaciones = correlaciones2(W,Y(:,1:3),pcuant,pnominal);

identif_cuantis(W(:,1:p1),Y);
identif_cualis(W(:,p1+1:p),Y);
```