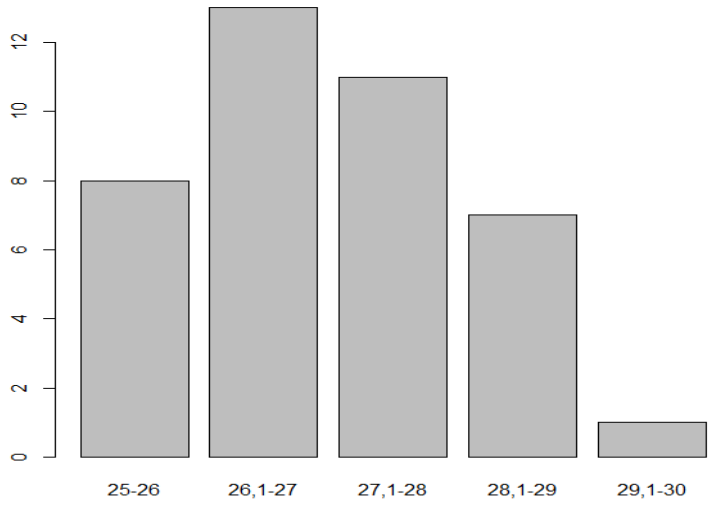
PRIMERA TAREA

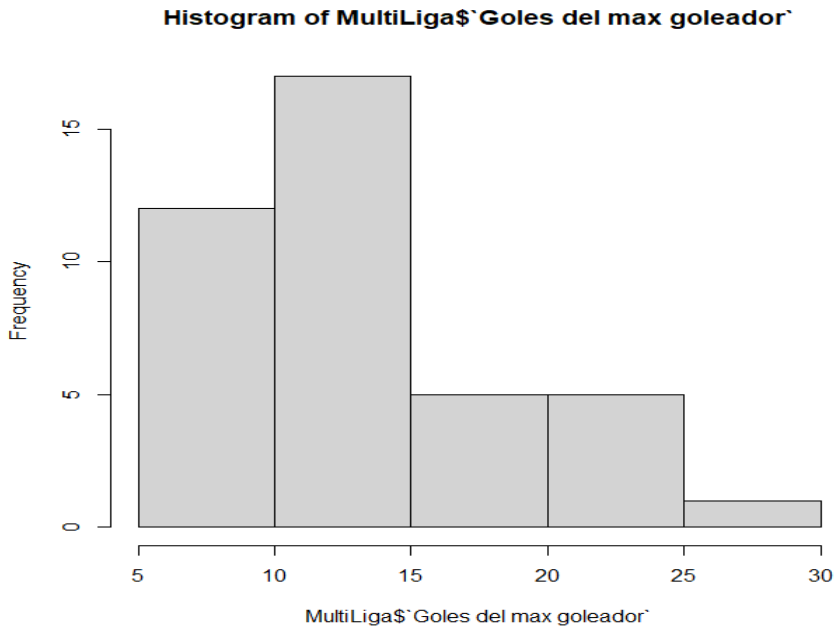
Daniel Aramburu y Guillermo Palomo Grupo 28

1. Descripción de variables

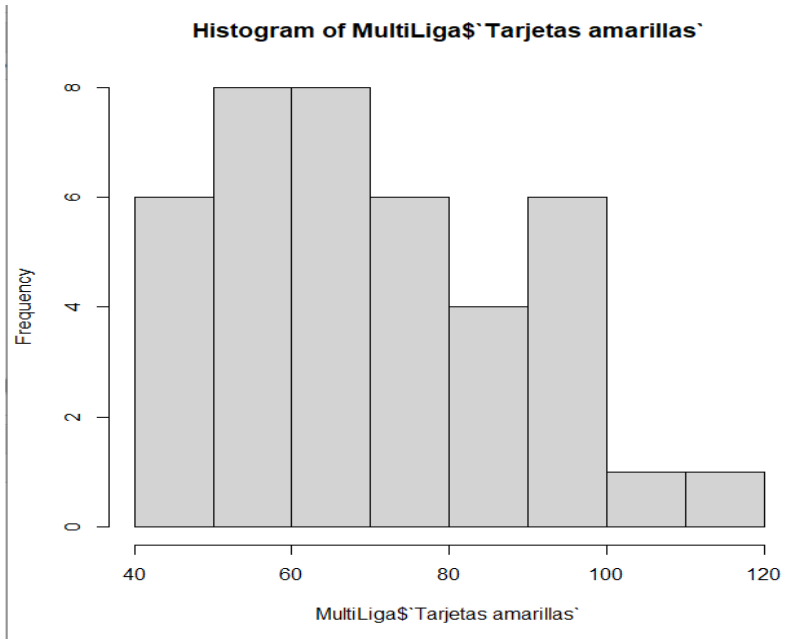
- Equipo: variable cualitativa que identifica el nombre de los equipos de la Liga Santander y los de la Premier League. Son 20 equipos por cada liga.
- Puesto: variable cualitativa que establece la posición de cada equipo en la clasificación de su liga.
- Rango de edad media: variable cualitativa categórica multi-estado que muestra la edad media de cada club. Como podemos observar en el gráfico a continuación, el rango de edad media en el que hay más jugadores es el de 26 a 27 años.



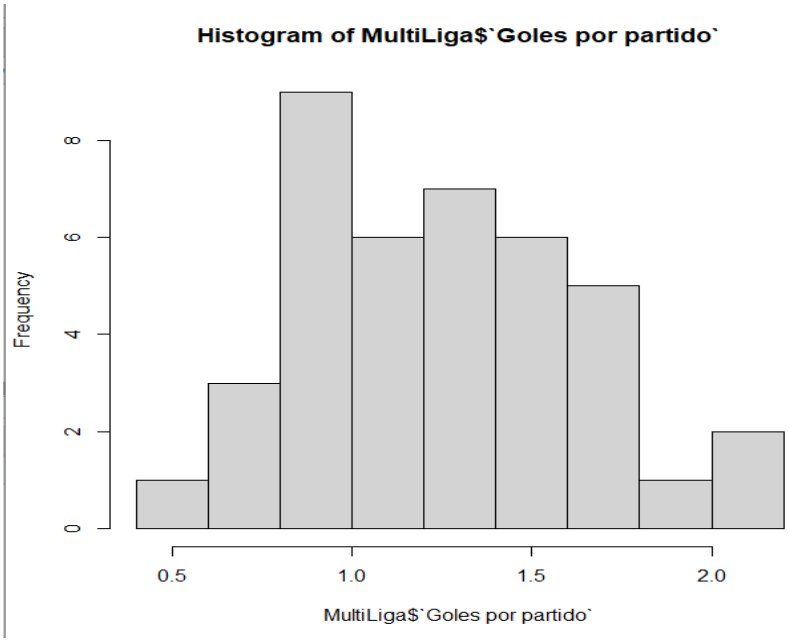
- Puestos importantes: variable cualitativa categórica multi-estado que muestra si un equipo está en posiciones importantes de la clasificación, es decir, si está en descenso, media tabla, Europa League, o Champions.
- Liga: es una variable cualitativa binaria que muestra un 0 si pertenece el equipo a la Liga española, y un 1 si pertenece a la inglesa.
- Recién ascendido: es una variable cualitativa binaria que muestra un 0 si el equipo no ha ascendido recientemente, y un 1 si ha ascendido recientemente.
- Goles del máximo goleador: es una variable cuantitativa discreta que ofrece la cantidad de goles que ha marcado el máximo goleador de cada equipo. En el siguiente gráfico podemos observar que la mayoría de los máximos goleadores de cada equipo han marcado entre 10 y 15 goles.



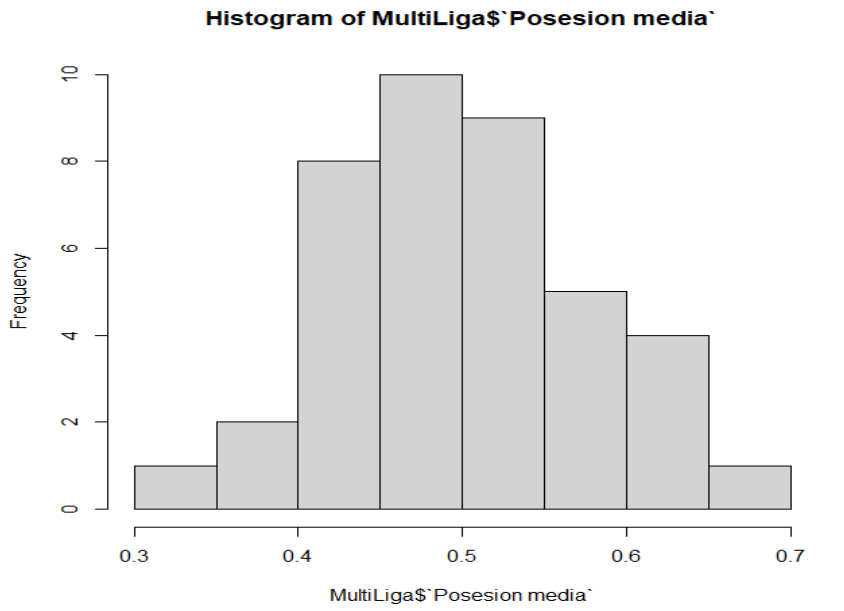
- Tarjetas amarillas: variable cuantitativa discreta que aporta el número de tarjetas amarillas recibidas en cada equipo. En el gráfico a continuación podemos observar que la gran mayoría de equipos reciben entre 50 y 70 amarillas por temporada.



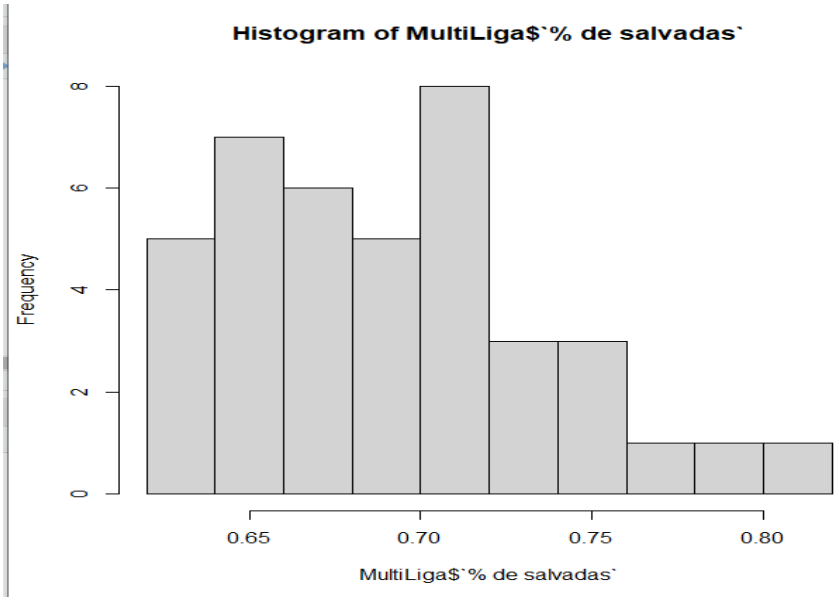
- Goles por partido: variable cuantitativa continua que muestra la media de goles que mete un equipo por partido. En el siguiente gráfico podemos observar que la mayoría de los goles por partido está entre 0.75 y 1.75.



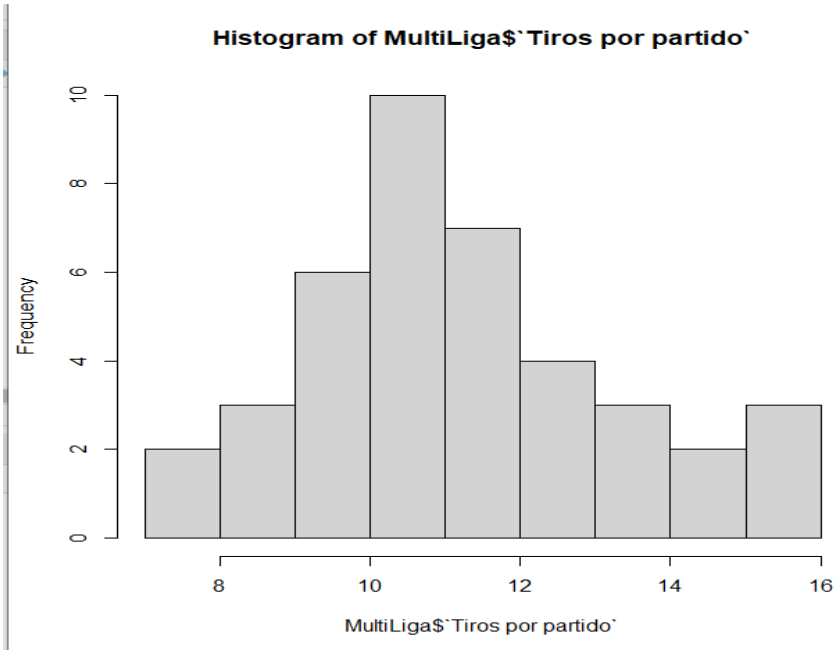
- Posesión media: variable cuantitativa continua que aporta la posesión media de cada equipo por partido. En el gráfico podemos ver que la mayoría de equipos tienen una posesión media entre 40% y 55%.



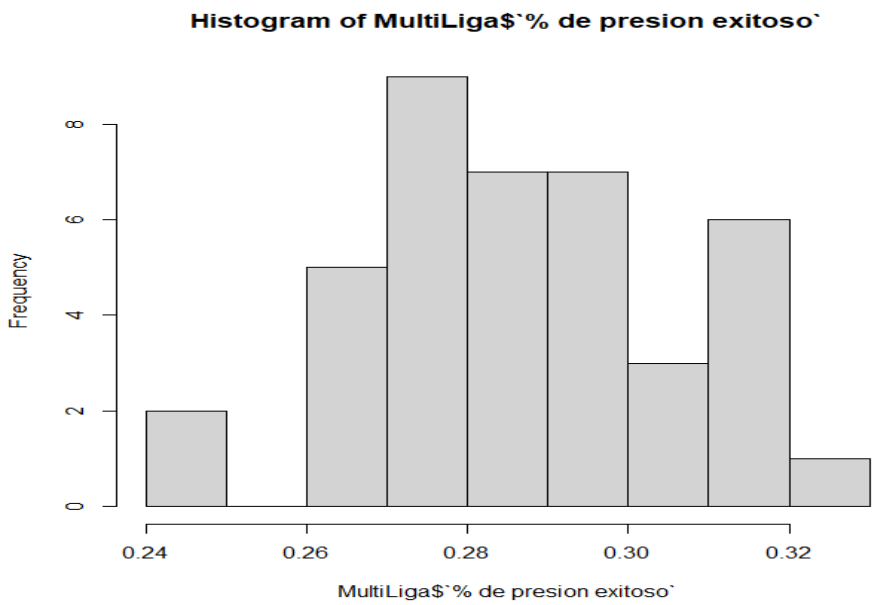
- % de salvadas: variable cuantitativa continua que muestra el porcentaje medio de paradas que tienen los porteros de cada equipo. El gráfico muestra que en la mayoría de los equipos, el porcentaje de salvadas oscila entre 60% y 70% aproximadamente.



- Tiros por partido: variable cuantitativa continua que ofrece la media de tiros que realiza cada equipo. En el gráfico podemos observar que la media de tiros por partido de los equipos está entre 9 y 12 tiros.



- % de presión exitoso: variable cuantitativa continua que muestra el porcentaje medio de presión efectiva realizada por cada equipo. En el gráfico podemos ver que el porcentaje exitoso medio de los equipos oscila entre 26% y 30%.



2. Ejercicios y conclusiones

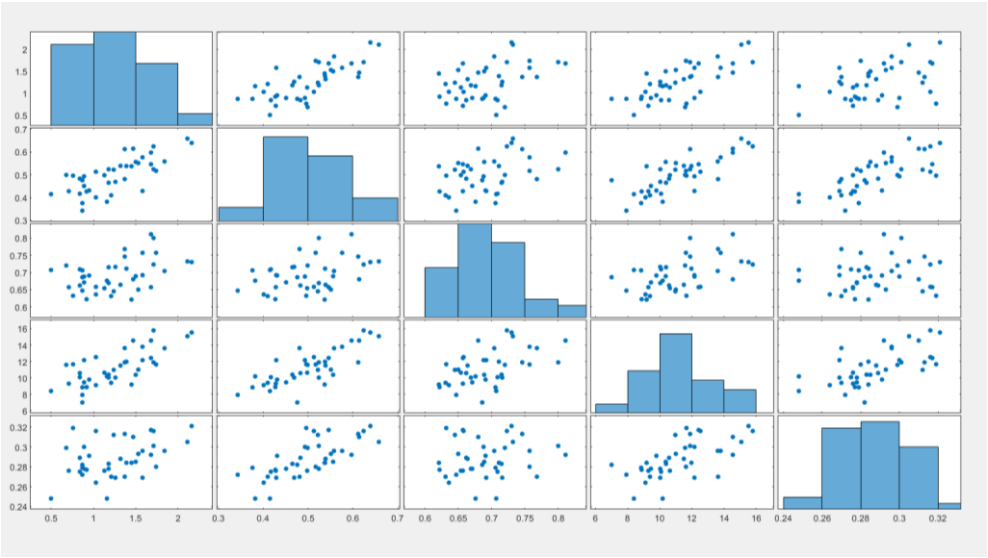
2.1. Variables cuantitativas:

En primer lugar, a partir de nuestros datos, que hemos llamado con la variable Y, calculamos: la matriz de covarianzas SY (función cov(Y)); matriz de correlaciones RY (función corr(Y)); un vector de medias MY (función mean (Y)); y por último, el gráfico de dispersión matricial (función plotmatrix(Y)).

	1	2	3	4	5
1	0.1596	0.0225	0.0082	0.6204	0.0031
2	0.0225	0.0059	0.0014	0.1295	0.0010
3	0.0082	0.0014	0.0022	0.0517	1.0720e-04
4	0.6204	0.1295	0.0517	4.5605	0.0252
5	0.0031	0.0010	1.0720e-04	0.0252	3.4546e-04

	1	2	3	4	5
1	1	0.7323	0.4383	0.7271	0.4111
2	0.7323	1	0.3892	0.7896	0.7294
3	0.4383	0.3892	1	0.5194	0.1238
4	0.7271	0.7896	0.5194	1	0.6358
5	0.4111	0.7294	0.1238	0.6358	1
6					

	1	2	3	4	5
1	1.2550	0.5004	0.6923	11.2353	0.2880



Además, comprobamos si las variables de nuestros datos se distribuyen normalmente o no, para en caso negativo, realizar transformaciones no lineales.

Comprobamos usando R y mediante el shapiro.test, que todas las variables son normales, aunque la que menos se podría parecer a una normal es la variable de porcentaje de salvadas (Variable 3 en el código). Es por ello que vamos a aplicar las transformaciones no lineales en esta variable, aunque no tendríamos por qué hacerlo ya que realmente es una normal y no haría falta.

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

        Shapiro-Wilk normality test

data:  QMulti[, 1]
W = 0.97301, p-value = 0.4458

> shapiro.test(QMulti[,2])

        Shapiro-Wilk normality test

data:  QMulti[, 2]
W = 0.98492, p-value = 0.8619

> shapiro.test(QMulti[,3])

        Shapiro-Wilk normality test

data:  QMulti[, 3]
W = 0.96095, p-value = 0.1805
```

```
> shapiro.test(QMulti[,4])

        Shapiro-Wilk normality test

data:  QMulti[, 4]
W = 0.97182, p-value = 0.4101

> shapiro.test(QMulti[,5])

        Shapiro-Wilk normality test

data:  QMulti[, 5]
W = 0.96502, p-value = 0.2477
```

Para esta variable realizamos el logaritmo (función log). Para estos nuevos datos con las transformaciones vamos a emplear la variable X.

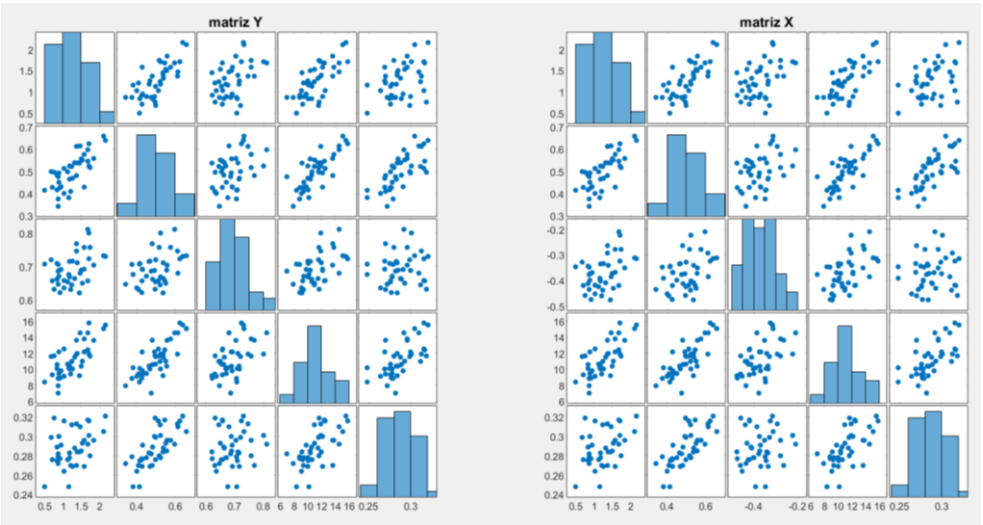
```
> shapiro.test(QMulti[,3])

        Shapiro-Wilk normality test

data:  QMulti[, 3]
W = 0.97084, p-value = 0.3826
```

Una vez realizado todo ello vamos a calcular: su matriz de covarianzas SX (función cov(X)); matriz de correlaciones SY (función corr(X)); un vector de medias MX (función mean (X)); y por último, el gráfico de dispersión matricial (función plotmatrix(X)).

	1	2	3	4	5
1	0.1596	0.0225	0.0115	0.6204	0.0031
2	0.0225	0.0059	0.0020	0.1295	0.0010
3	0.0115	0.0020	0.0044	0.0735	1.5016e-04
4	0.6204	0.1295	0.0735	4.5605	0.0252
5	0.0031	0.0010	1.5016e-04	0.0252	3.4546e-04

	1	2	3	4	5
1	1	0.7323	0.4336	0.7271	0.4111
2	0.7323	1	0.3881	0.7896	0.7294
3	0.4336	0.3881	1	0.5196	0.1220
4	0.7271	0.7896	0.5196	1	0.6358
5	0.4111	0.7294	0.1220	0.6358	1

	1	2	3	4	5
1	1.2550	0.5004	-0.3700	11.2353	0.2880



Vemos cómo cambia ligeramente la variable 3 de porcentaje de salvadas de porteros.

A partir del gráfico podemos concluir la relación entre las distintas variables.

Por un lado, la variable 1 y la 2 (goles por partido y posesión media); la 1 y la 4 (goles por partido y tiros por partido); la 2 y la 4 (posesión media y tiros por partido); y por último, la 2 y la 5 (posesión media y porcentaje de presión exitoso) tienen una relación lineal positiva.

Por otro lado, la variable 1 y la 5 (goles por partido y porcentaje de presión exitoso); las variables 4 y 5 (tiros por partido y porcentaje de presión exitoso); y por último, la variable 3 (porcentaje de salvadas) con el resto de variables, no están claramente correladas. Esto lo podemos comprobar con la dispersión de los puntos en el gráfico.

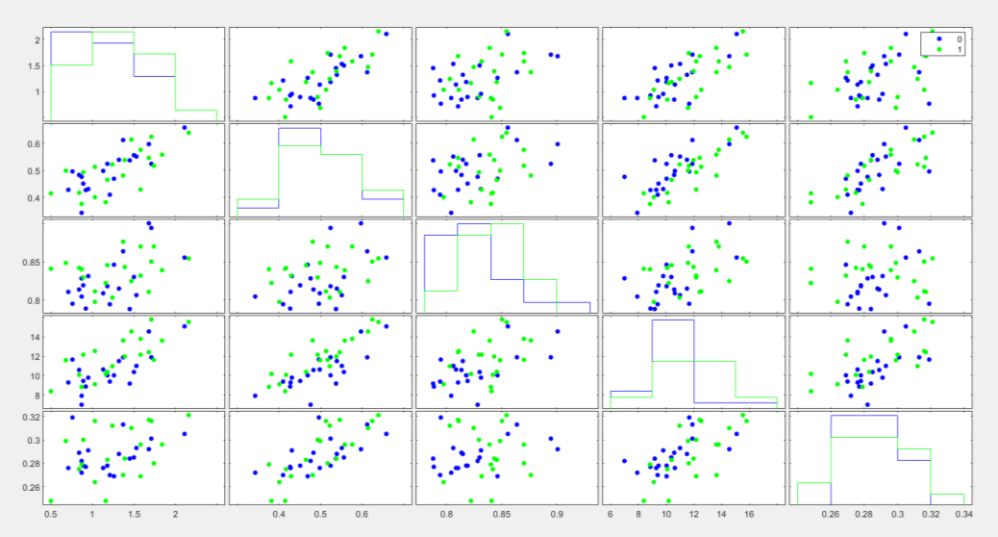
Lo último relacionado con las variables cuantitativas es obtener las medidas escalares de dispersión.

Para los datos originales, calculamos la variación total (VTY= trace(SY)) = 4.7286, la variación generalizada (VGY = det(SY)) = 1.2875e-10, y la η2, mediante el código 1-det(RY) = 0.9600.

Para los datos transformados realizamos el mismo código sustituyendo la Y de los datos originales por la X, calculamos la variación total (VTX= trace(SX)) = 4.7308, la variación generalizada (VGX= det(SX)) = 2.5984e-10, y la η2, mediante el código 1-det(RX) = 0.9601.

2.2. Variable binaria:

Creamos el gráfico de dispersión matricial con gplotmatrix separando según la variable Liga los equipos que son de LaLiga (0) y de Premier (1).



En este gráfico podemos apreciar las mismas relaciones entre variables mencionadas anteriormente, con la diferencia de que este gráfico muestra mediante dos colores diferentes las dos ligas estudiadas. Es decir, las relaciones entre variables se mantienen en las dos ligas.

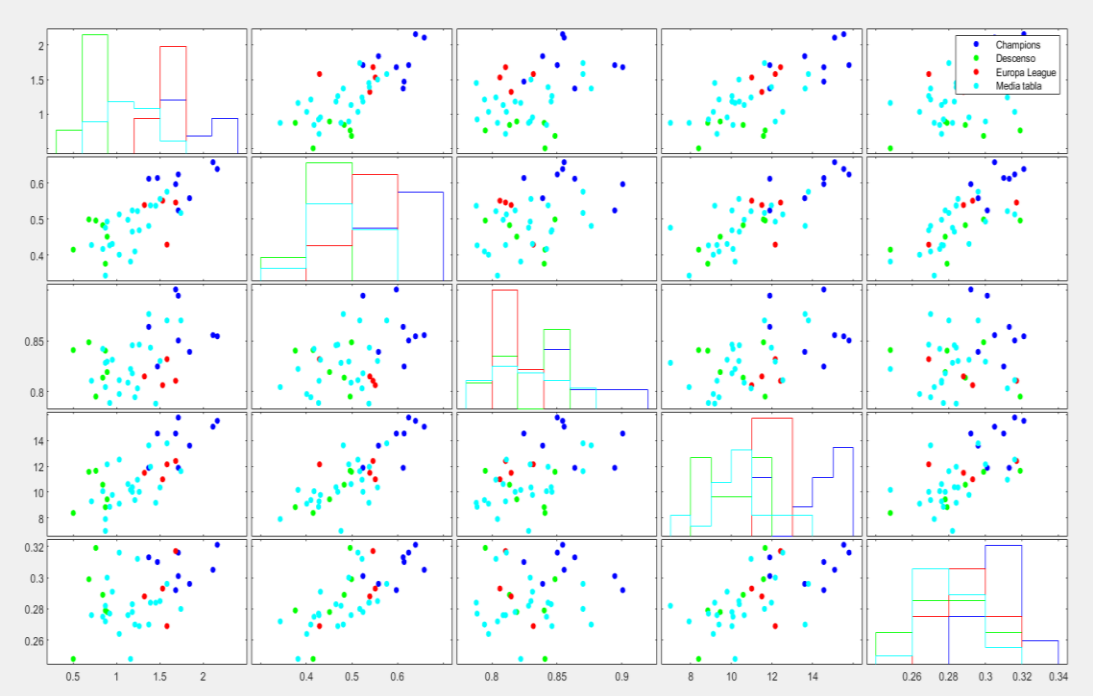
Ahora para las dos ligas vamos a hacer un contraste de comparación de medias con el estadístico T2 de Hotelling:

```
T2 =

   18.2784

>> F = (nLL+nP-p-1)/((nLL+nP)*p)*T2

F =

    3.1073

>> percentil = finv(0.95,p,nLL+nP-1)

percentil =

    2.4558

>> p_valorT2 = 1-fcdf(F,p,nLL+nP-p-1)

p_valorT2 =

    0.0204
```

Como podemos ver por el p-valor de 0.0204 para una significación de 0.05 se rechaza la hipótesis nula de igualdad de medias por bastante poco, ya que para un nivel de significación de 0.01 sí que concluiríamos que las medias son iguales. En definitiva, rechazamos la hipótesis de igualdad de medias.

2.3. Variable categórica multi-estado:

Para la variable categórica multi-estado hemos elegido los puestos importantes de la clasificación siendo estos Champions, Europa League, media tabla y descenso.



Esta vez sí se nota que los grupos son bastante diferentes entre ellos lo cual tiene sentido ya que los mejores equipos en estadísticas suelen quedar mejor en la clasificación.

Para esta variable de puestos importantes vamos a hacer un contraste de comparación de medias mediante la Lambda de Wilks:

```
>> Lambda = det(W)/det(T)

Lambda =

    0.2198

>> [F,m,n] = wilkstof(Lambda, 5, 36, 3)

F =

    4.3390


m =

    15


n =

    89

>> p_valorW=1-fcdf(F,m,n)

p_valorW =

   5.2312e-06
```

Por lo ínfimo que es el p-valor del contraste con 5.2312e-6, podemos rechazar la hipótesis nula de igualdad de medias lo cual era esperable ya que los grupos son bastante diferenciados entre ellos ya que un equipo en descenso no suele poder competir contra uno que acabe la temporada en puestos de Champions.