Análisis de Componentes Principales Premier League 19-20

Gómez Jiménez Aaron Mauricio

2023-04-09

Con el objetivo de identificar las variables que son de mayor importancia para ser campeón en la Premier League, realizaremos un análisis de componentes principales, como marco referencial el campeón de esta liga es el equipo que más partidos ganados tenga a lo largo de los 38 partidos de la temporada.

Realizando un análsis exploratorio de estadísticas básicas obtenemos los siguientes resultados

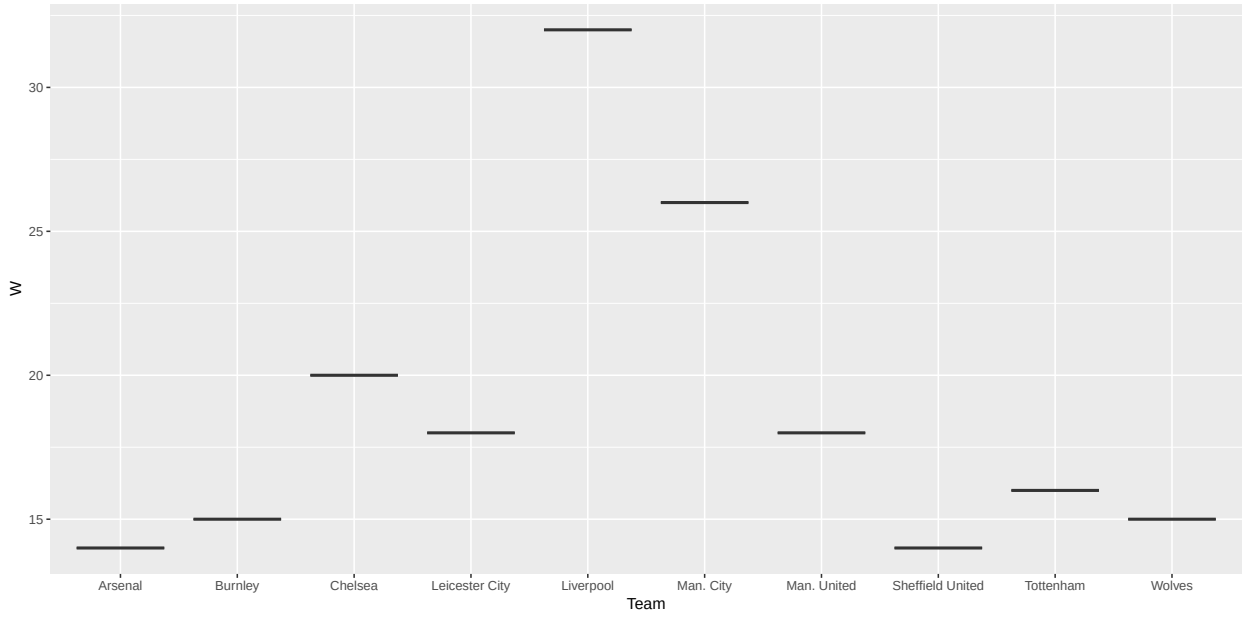
```
resumen<-describe(datos, fast=TRUE)
resumen
```

```
##      vars  n mean    sd median min  max range  skew kurtosis   se
## Team    1 20  NaN    NA     NA Inf -Inf  -Inf    NA       NA   NA
## W       2 20 14.4  6.35   14.0   5   32    27  1.08     0.93 1.42
## D       3 20  9.2  3.27    9.5   3   14    11 -0.25    -0.83 0.73
## L       4 20 14.4  5.63   14.5   3   27    24  0.22    -0.38 1.26
## GF      5 20 51.7 18.81   46.5  26  102    76  1.02     0.45 4.21
## GA      6 20 51.7 11.87   52.0  33   75    42  0.07    -1.14 2.65
## GD      7 20  0.0 28.37   -8.0 -49   67   116  0.65    -0.18 6.34
```

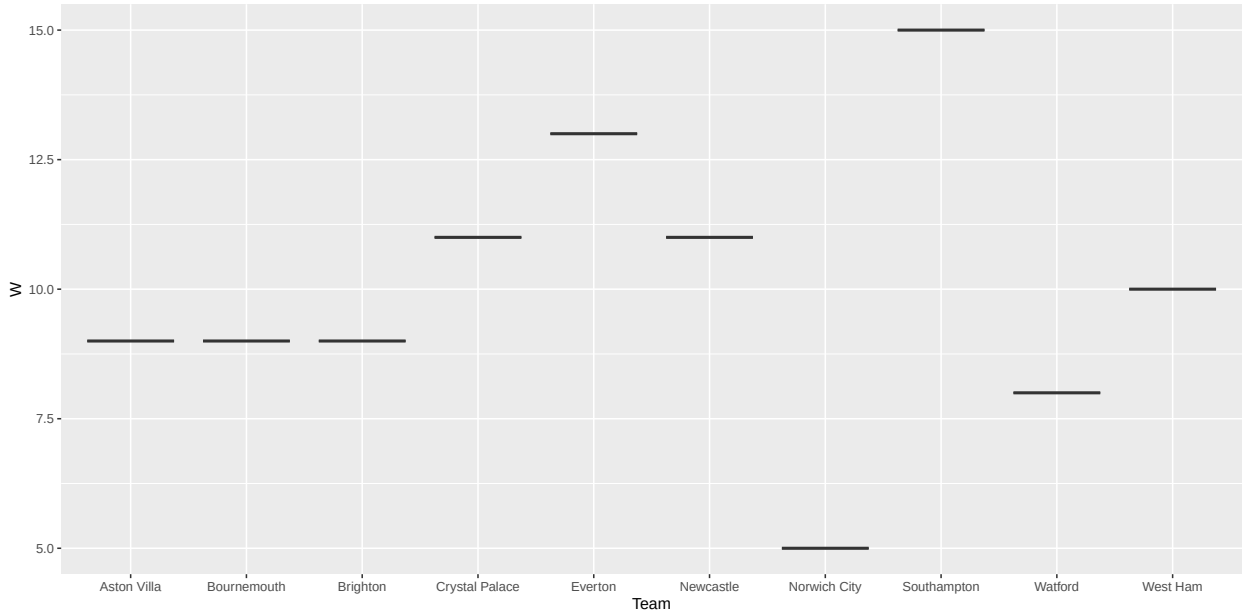
Donde W es el número de partidos ganados, D partidos empatados, L partidos perdidos, GF goles a favor, GA goles en contra y GD la diferencia de goles a favor menos goles en contra.

Partidos Ganados

```
ganados_1<-ggplot(datos[1:10,], aes(x = Team, y = W, fill=Team)) +
  geom_boxplot() + theme(legend.position = "none")
ganados_1
```



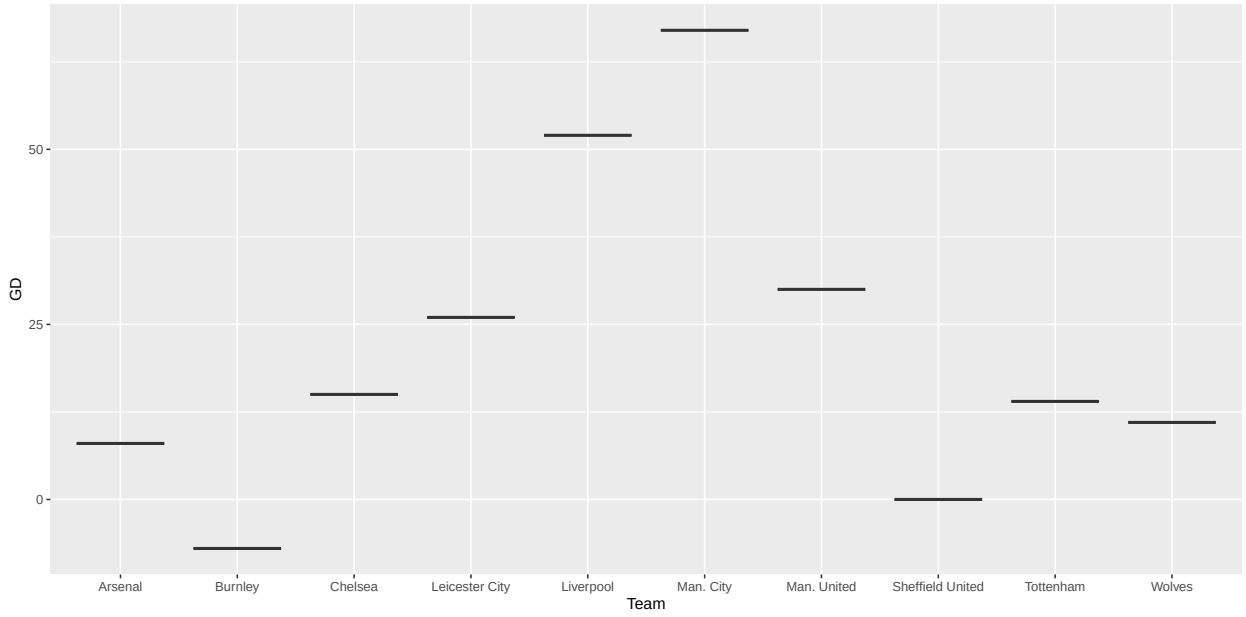

```
ganados_1.1<-ggplot(datos[11:20,], aes(x = Team, y = W, fill=Team)) +
  geom_boxplot() + theme(legend.position = "none")
ganados_1.1
```



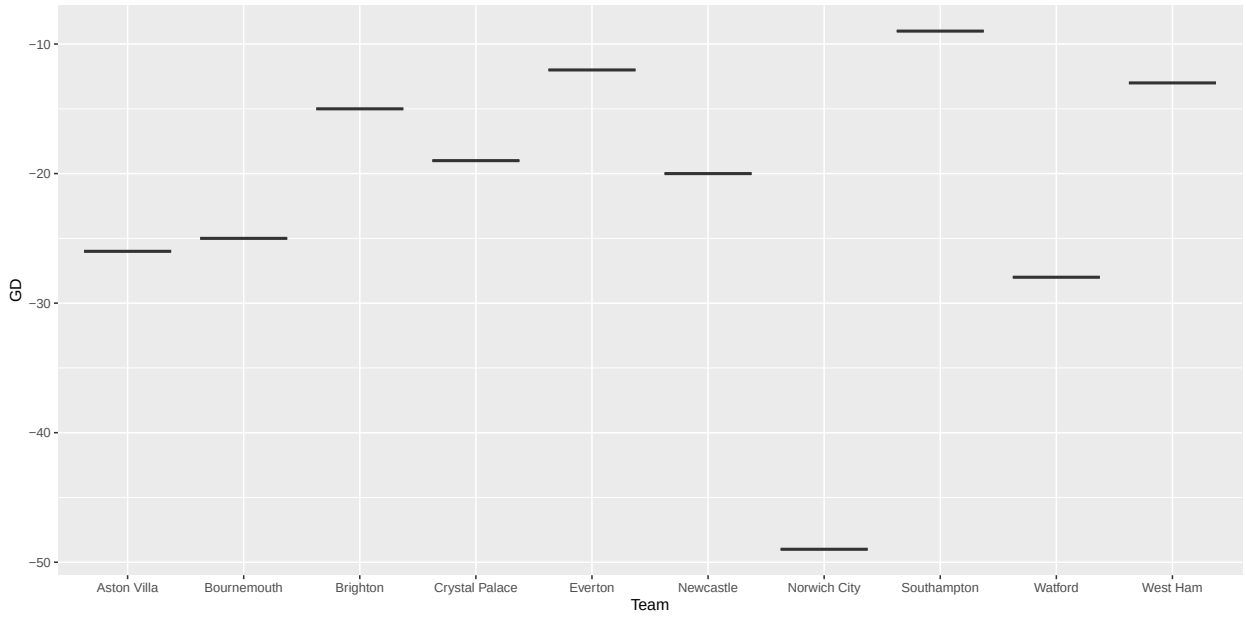
El equipo que obtuvo mayor número de victorias fue el Liverpool, seguido del Manchester City, y los que menos partidos ganaron fueron Norwich City y Watford, el promedio de juegos ganados fue de 15 partidos con 52 goles en promedio por equipo para toda la temporada, con una media de 27 goles por jornada, es decir casi 3 goles por partido en promedio.

Diferencia de goles

```
diferencia_1<-ggplot(datos[1:10,], aes(x = Team, y = GD, fill=Team)) +
  geom_boxplot() + theme(legend.position = "none")
diferencia_1
```




```
diferencia_1.1<-ggplot(datos[11:20,], aes(x = Team, y = GD, fill=Team)) +
  geom_boxplot() + theme(legend.position = "none")
diferencia_1.1
```



El equipo que mejor diferencia de goles tiene fue el Manchester City seguido del campeón de esa temporada, el Liverpool, y los que peor diferencia de goles tienen son nuevamente el Norwich City y Watford.

Análisis de Componentes Principales

Para realizar el análisis de componentes principales es necesario tener la matriz de datos, tambien es recomendable estandarizar los datos.

Haciendo el PCA centrado y escalado obtenemos

```
PCA_centrado<-prcomp(datos, center=TRUE, scale=TRUE)
```

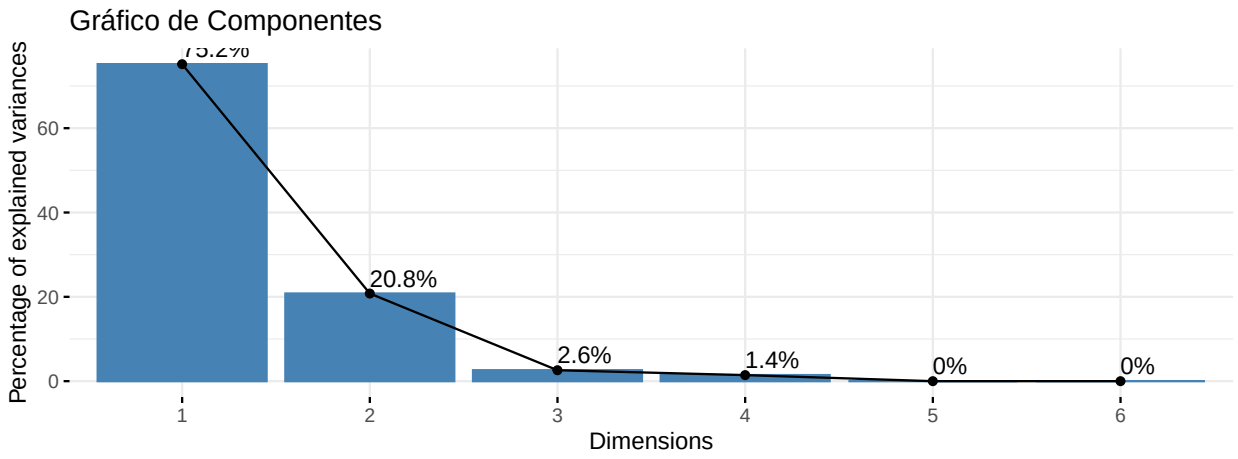
Primero revisaremos los eigenvalues resultantes del PCA para ver la proporción de varianza explicada por cada componente y decidir con que número de componentes nos quedamos, utilizaremos la regla de Kaiser, es decir, eigenvalues mayores a 1, además nuscamos una varianza acumulada mayor al 90%.

```
eigenvalues<-get_eigenvalue(PCA_centrado)
eigenvalues
```

```
##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1 4.510922e+00     7.518203e+01                     75.18203
## Dim.2 1.247250e+00     2.078751e+01                     95.96954
## Dim.3 1.555633e-01     2.592722e+00                     98.56226
## Dim.4 8.626440e-02     1.437740e+00                    100.00000
## Dim.5 1.577702e-32     2.629504e-31                    100.00000
## Dim.6 3.471978e-33     5.786630e-32                    100.00000
```

Podemos notar que el número de componentes de acuerdo a los criterios es 2 componentes. Haciendo un análisis gráfico.

```
fviz_screeplot(PCA_centrado, main="Gráfico de Componentes", addlabels=TRUE)
```



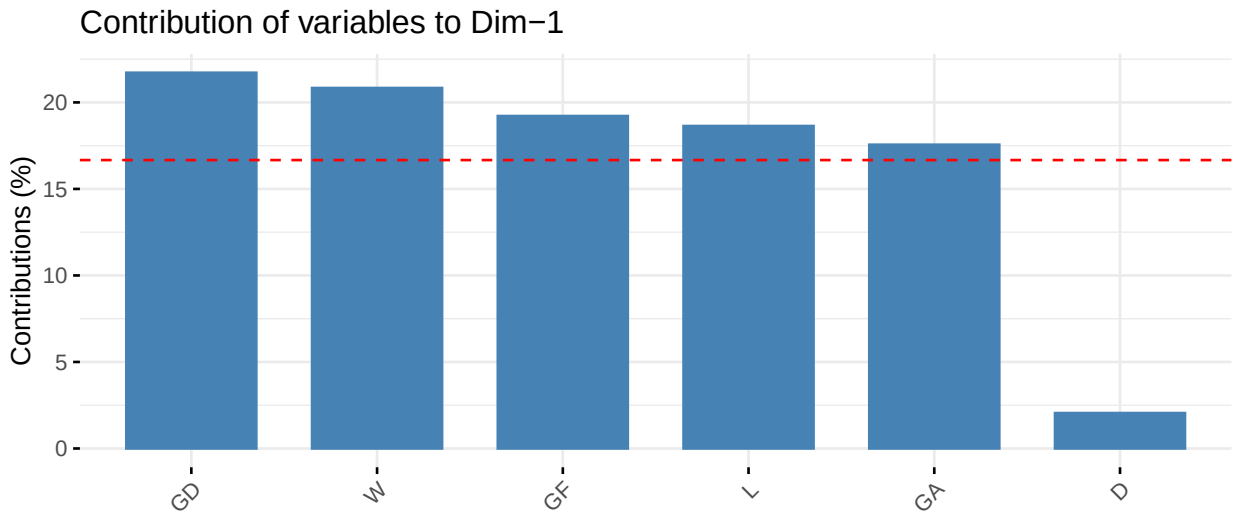
Ahora nos centraremos en la contribución de las variables a cada componente, poniendo especial atención en los dos primeros componentes

```
varianza<-get_pca_var(PCA_centrado)
head(varianza$contrib)
```

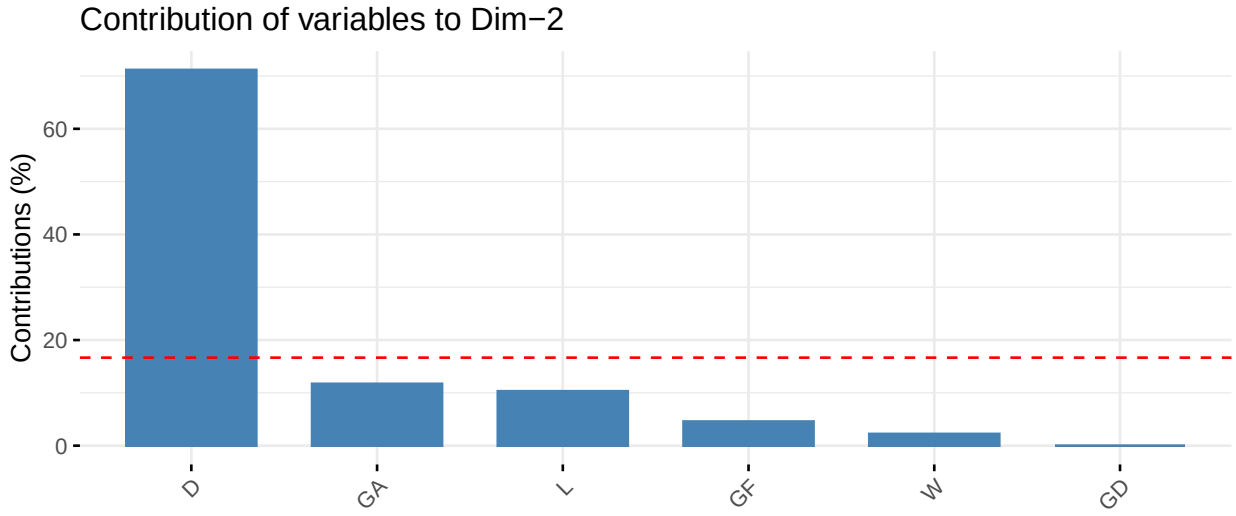
```
##        Dim.1        Dim.2     Dim.3       Dim.4     Dim.5     Dim.6
## W  20.832188 2.233041e+00 11.709459 16.47149761 36.351768 12.402047
```

```
## D   2.047506 7.115392e+01 11.830580  2.04835796  9.633124  3.286510
## L  18.636653 1.031761e+01  3.468631 29.25055772 28.577003  9.749548
## GF 19.213516 4.579664e+00 48.947945  0.03290859  6.925770 20.300196
## GA 17.555001 1.171558e+01 14.903355 44.97893103  2.759305  8.087826
## GD 21.715136 1.839042e-04  9.140030  7.21774708 15.753031 46.173872
```

```
par(mfrow = c(1, 2))
fviz_contrib(PCA_centrado, choice = "var", axes = 1, top = 6)
```

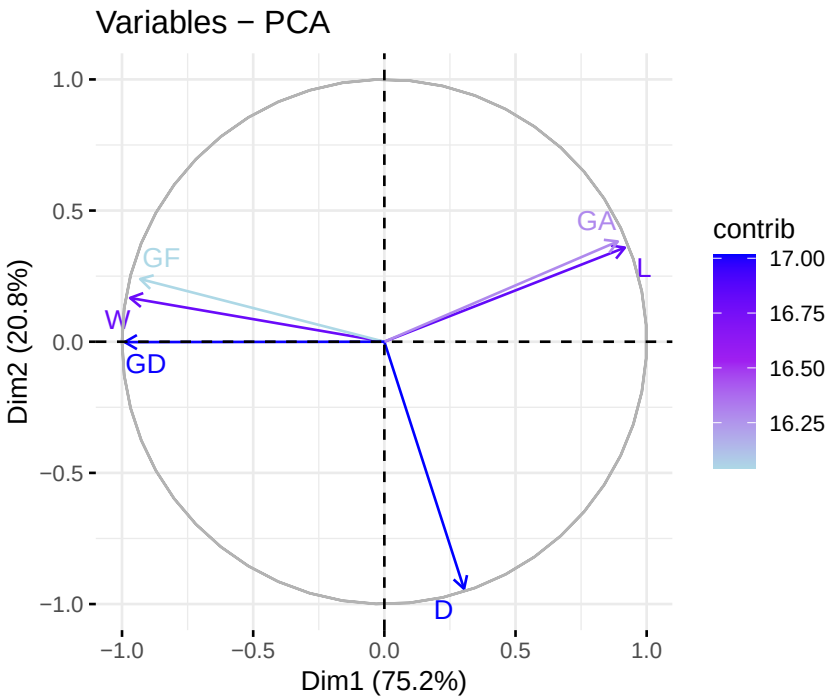



```
fviz_contrib(PCA_centrado, choice = "var", axes = 2, top = 6)
```



Podemos notar que para el primer componente 5 variables superan el 16.6, es decir estan sobre la media esperada sobre la contribución de cada variable, para el segundo componente solo la variable D tiene contribución muy significativa.

```
fviz_pca_var(PCA_centrado, col.var = "contrib", repel =TRUE,gradient.cols = c("lightblue", "purple", "bl
```



Como podemos observar lo que mas valor le da al primer componente es el numero partidos perdidos y goles en contra es decir cosas negativas en el sentido futbolistico, inversamente los partidos ganados y goles a favor los pondera por igual, para el segundo componente observamos que el mas significativo es el numero de empates, asi concluimos que entre mas a la izquierda y arriba mejor es la posición en la tabla del equipo.

Obteniendo las coordenadas

```
PCA_centrado$rotation
```

```
##           PC1          PC2        PC3         PC4        PC5        PC6
## W  -0.4564229  0.149433637 -0.3421909 -0.40585093 -0.6029243  0.3521654
## D   0.1430911 -0.843527843  0.3439561 -0.14312086 -0.3103727  0.1812873
## L   0.4317019  0.321210308  0.1862426  0.54083785 -0.5345746  0.3122427
## GF -0.4383322  0.214001505  0.6996281 -0.01814073  0.2631686  0.4505574
## GA  0.4189869  0.342280325  0.3860486 -0.67066334 -0.1661115 -0.2843910
## GD -0.4659950 -0.001356113  0.3023248  0.26865865 -0.3969009 -0.6795136
```

Para el primer componente notamos que los partidos ganados w, los goles a favor GF, y la diferencia de goles GD tienen la misma dirección y los pondera por igual en el primer componente, de esta manera sospechamos que hay algun tipo de cluster estre las variables con signo positivo y negativo. Para el segundo componente notamos que los partidos empatado D, tienen un gran peso siendo la unica significativa con signo negativo.

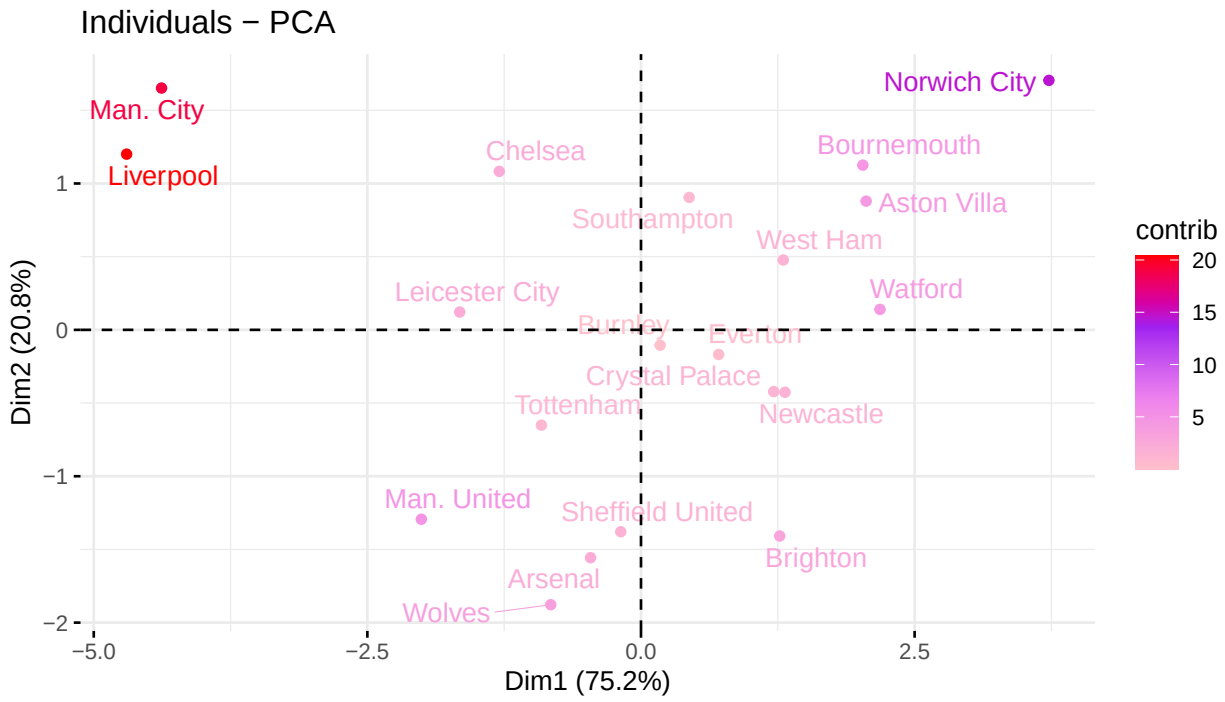
Análisis Individual

Podemos ver la influencia de cada observación en los componentes principales

```
ind<-get_pca_ind(PCA_centrado)
ind$contrib[1:10,1:2]
```

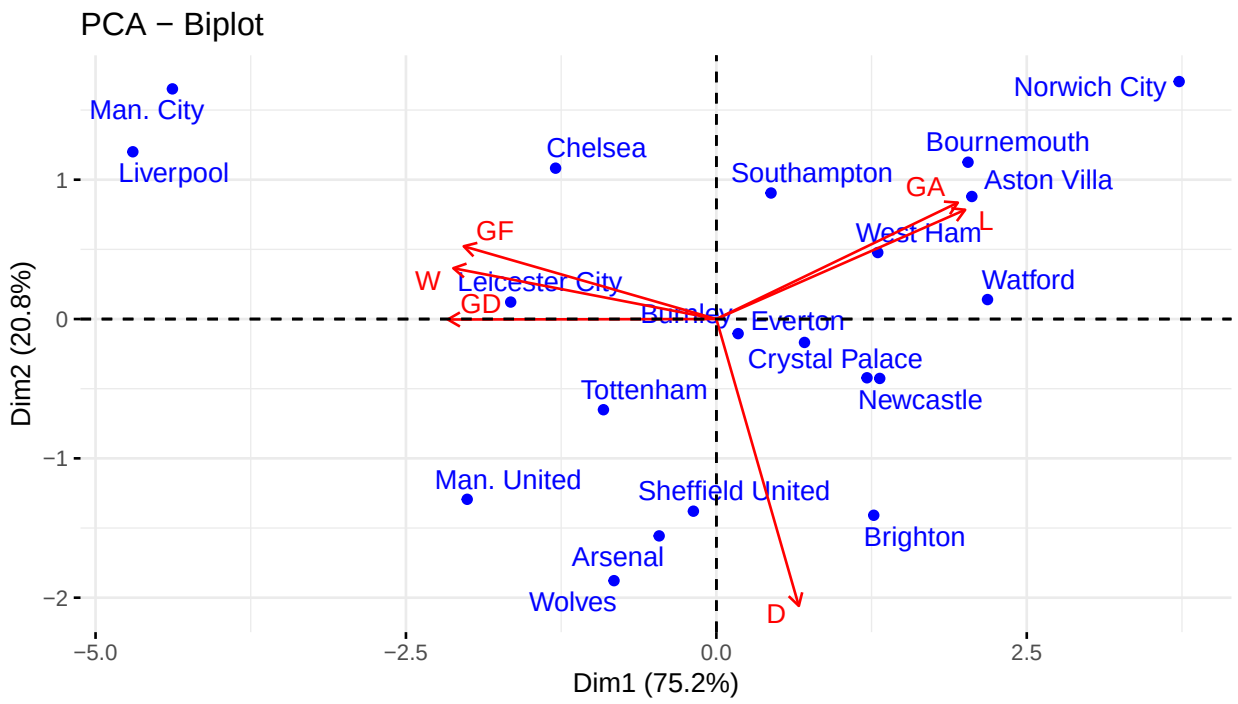
```
##                       Dim.1       Dim.2
## Liverpool        24.48130539  5.77623464
## Man. City        21.27110990 10.93715866
## Man. United       4.46368547  6.71124651
## Chelsea           1.85795011  4.69911149
## Leicester City    3.04093578  0.05936932
## Tottenham         0.91656582  1.69843025
## Wolves            0.75302446 14.13451910
## Arsenal           0.23521841  9.71283603
## Sheffield United  0.03790371  7.62470827
## Burnley           0.03403096  0.04400170
```

```
fviz_pca_ind(PCA_centrado, col.ind = "contrib",
             gradient.cols = c("pink", "violet","purple" , "red"),
             col.var="green",
             repel = TRUE
             )
```



Aqui podemos observar que aunque el Arsenal, Wolves y Brighton tuvieron el mismo numero de empates, lo que les cambio la ubicación en el plot es que wolves gano 15 partidos, mientras Arsenal 14 y Brighton 9.

```
fviz_pca_biplot(PCA_centrado, repel = TRUE,
                col.var = "red",
                col.ind = "blue"
                )
```



Conclusiones

Podemos concluir que bastan 2 componentes para explicar el 96% de la varianza de los datos, las varaibles que influyen en mayor cantidad para el primer componente son GD, W, GF, L , GA en ese orden y para el segundo componente la variable D influye en un 85% siendo la única significativa para el segundo componente, de forma que en el biplot entre más a la izquierda y arriba este los equipos mayor es la cantidad de partidos gandos, entre más arriba mayor es el número de goles que tuvieron a su favor, en contraparte entre más abajo este ubicados los equipos mayor es el número de partidos perdidos, entre más a la izquierda su diferencia de goles es cercana a cero o negativa.

Es importante notar que para la primer componente la diferencia de goles es la que mayor contribución tiene, seguida de los partidos ganados, lo cual puede parecer contraintuitivo, sin embargo estas dos variable y los goles a favor van en la misma dirección lo cual es lógico ya que son variables que consideramos positivas para nuestro objetivo, y para la segunda componente vemos que el vector de los partidos empatados esta a 90 grados con la variable GD lo cual no quiere indicar que no existe relación entre las variables, pero la variable D es la que mayor peso tiene en la segunda componente.