SESIÓN 4: ANALISIS DE COMPONENTES PRINCIPALES

Disponemos de una muestra de 2000 clientes de una entidad bancaria. Los datos se encuentran en el fichero “churn.txt”.

El objetivo es obtener una tipología de clientes según su posición bancaria. La posición bancaria de un cliente viene definida por el saldo en cada uno de los productos de pasivo y activo. En nuestro caso esta información viene reflejada de forma agregada en las siguientes variables:

```
[13] "Total_activo"          "Total_Plazo"          "Total_Inversion"
[16] "Total_Seguros"         "Total_Vista"
```

Para ello realizaremos primero un Análisis de Componentes Principales, para ver cuáles son los factores latentes que estructuran los datos y minimizar la parte de fluctuación aleatoria para a continuación efectuar el “Clustering” (en la 5ª sesión).

1. Lea el fichero “churn.txt”. Razone si realizar un ACP con datos estandarizados o sin estandarizar y efectúe un Análisis de Componentes Principales como activas las variables de posición antes especificadas.

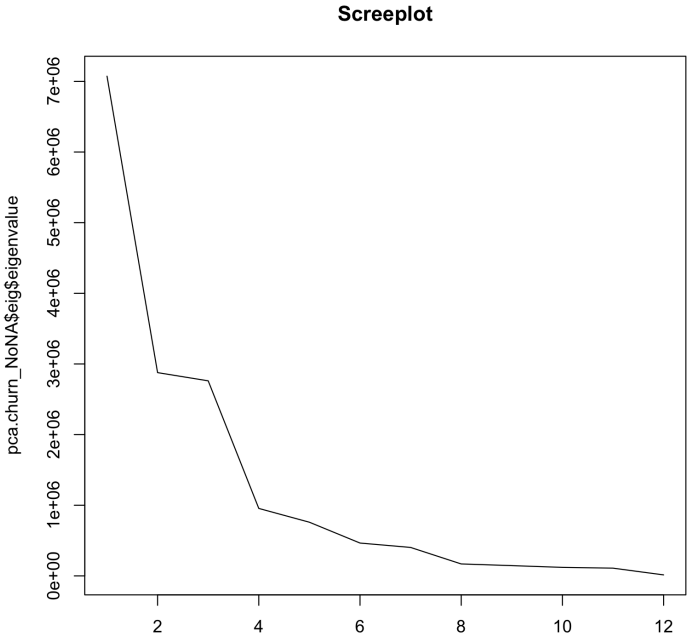
Haré un ACP con datos no estandarizados, dado que todas las unidades de las variables continuas son las mismas. La única que era una variable continua y la paso a categórica es la antigüedad. Que de todas formas al ser un número muy finito de valores, se puede hacer.

```
> churn_NoNA <- churn[complete.cases(churn),]
> churn_NoNA$antig <- as.factor(churn_NoNA$antig)
> pca.churn <-
PCA(churn_NoNA,quali.sup=c(1:12,18),quanti.sup=c(13:17),graph=T,scale.
unit=F) > pca.churn$eig
eigenvalue percentage of variance cumulative percentage of variance
```

```
  eigenvalue percentage of variance cumulative percentage of variance
comp 1  7073622.66   44.61820805        44.61821
comp 2  2877410.47   18.14980886        62.76802
comp 3  2761175.35   17.41663391        80.18465
comp 4   954987.45    6.02376335        86.20841
comp 5   759712.62    4.79203056        91.00044
comp 6   465030.39    2.93326681        93.93371
comp 7   402877.02    2.54122275        96.47493
comp 8   169110.91    1.06669895        97.54163
comp 9   145509.44    0.91782820        98.45946
comp 10  120557.54    0.76043938        99.21990
comp 11  109566.92    0.69111397        99.91101
comp 12   14107.42    0.08898522       100.00000
```

2. Obtenga la representación gráfica del ”Screeplot” (diagrama de los “eigenvalues”) y a la vista de las correlaciones entre las variables originales y las componentes principales, decida el número de dimensiones significativas. ¿Cuál es el porcentaje de variancia retenido?.

```
> plot(pca.churn$eig$eigenvalue,type="l",main="Screeplot")
```



Según la regla del último codo debería escoger como variables más significativas las anteriores a la 4, es decir, las 3 primeras. Pero eso nos bajaría la varianza a un valor cercano a 80, así que como es mejor coger ruido que no perder significancia, escogeré los 5 primeros.

El porcentaje de varianza es del `91.00044`.

3. Efectúe una rotación "varimax" para hacer más evidente los factores latentes (intangibles) presentes en sus datos activos. ¿Cuáles son en este caso estos factores latentes?.

Una vez rotados los factores quedan así:

```
> pca.churn.rot <- varimax(pca.churn$var$cor[,1:nd])
> pca.churn.rot
$loadings

Loadings:

                     Dim.1  Dim.2  Dim.3
oper_caj_Libreta                   -0.166
oper_ven_Libreta                   -0.109
dif_CC
dif_Libreta                 0.106
dif_Plazo                   0.927  0.372
dif_Ahorro
dif_Largo_plazo
dif_Fondos_inv        0.998
dif_Seguros
dif_Planes_pension
dif_Hipoteca                0.391 -0.919
dif_Prest_personales

               Dim.1 Dim.2 Dim.3
SS loadings    1.006 1.033 1.036
```

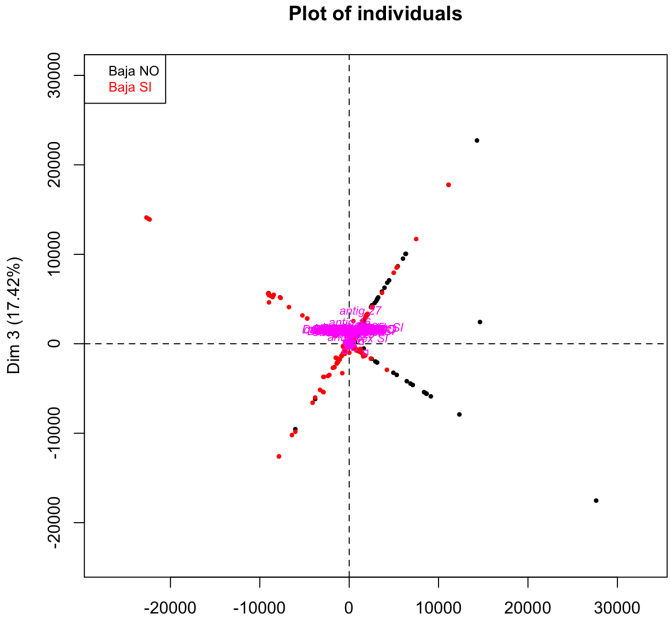
```
Proportion Var 0.084 0.086 0.086
Cumulative Var 0.084 0.170 0.256

$rotmat
             [,1]        [,2]        [,3]
[1,]  0.997492102 -0.05201745  0.04799677
[2,]  0.070181146  0.81478712 -0.57549679
[3,] -0.009171274  0.57742197  0.81639436
```

Las variables con más significancia, para la primera dimensión es la diferencia que ha habido durante el año en los fondos de inversión (`dif_Fondos_inv`), para la segunda dimensión la diferencia de los pagos que se tienen a plazos (`dif_Plazo`) y de forma negativa para la tercera dimensión, la diferencia de los pagos hechos por hipotecas (`dif_Hipoteca`).

4. Represente gráficamente la nube de puntos individuo, diferenciando los que han sido baja y los que no han sido baja. ¿Piensa Ud. que la *posición* de los clientes permitirá separar fácilmente la "baja" de la "no baja"?.

Después de probar con las diferentes posiciones posibles de los puntos, sí que parece ser, que según la visualización que he añadido más abajo, pueda haber cierto agrupamenti para las no-bajas en el cuadrante 1 y 4 y para las sí-bajas en el cuadrante 2 y 3.

```
> plot(pca.churn_NoNA, axes = c(2, 3), choix = c("ind"), habillage=1,
label="quali", title="Plot of individuals", cex=0.7)
```



5. Represente gráficamente el mapa de correlación de las variables activas. Sobre este mapa, represente la correlación de la variable "baja" considerada numérica (basta utilizar la función `as.numeric` de R) con las componentes principales. ¿Piensa Ud. que la variable "baja" esta correlacionada con las variables activas?

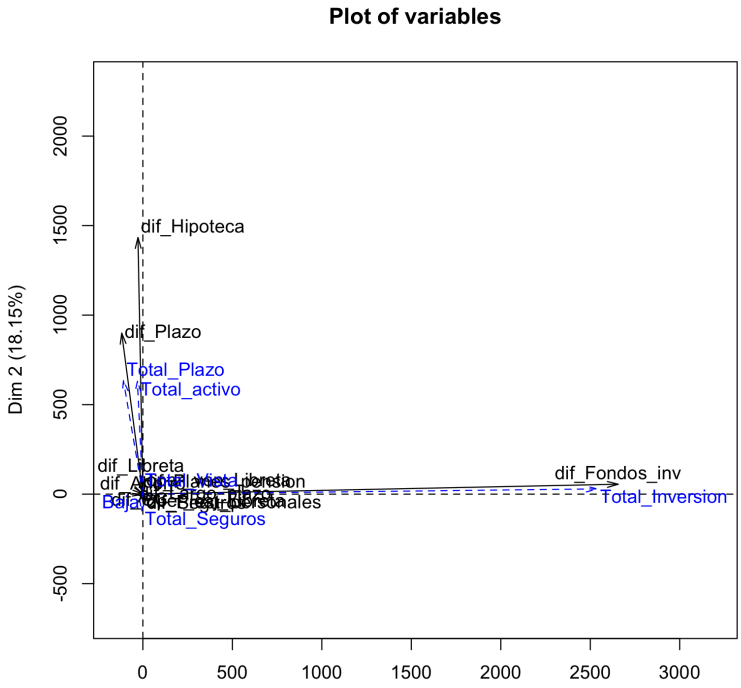
Siguiendo los mismos pasos que en el primer ejercicio, pero añadiendo las siguientes líneas:

```
> churn_NoNA$Baja <- as.numeric(churn_NoNA$Baja)
```

```
> summary(churn_NoNA)
```

Y enseñado el plot

```
> plot(pca.churn_NoNA, axes = c(1, 2), choix = c("var"), title="Plot
of variables")
```



Observamos que la variable Baja se encuentra precisamente en el origen de coordenadas. Por lo tanto podemos deducir que no tiene relación con ninguna de las otras variables activas.