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"
```

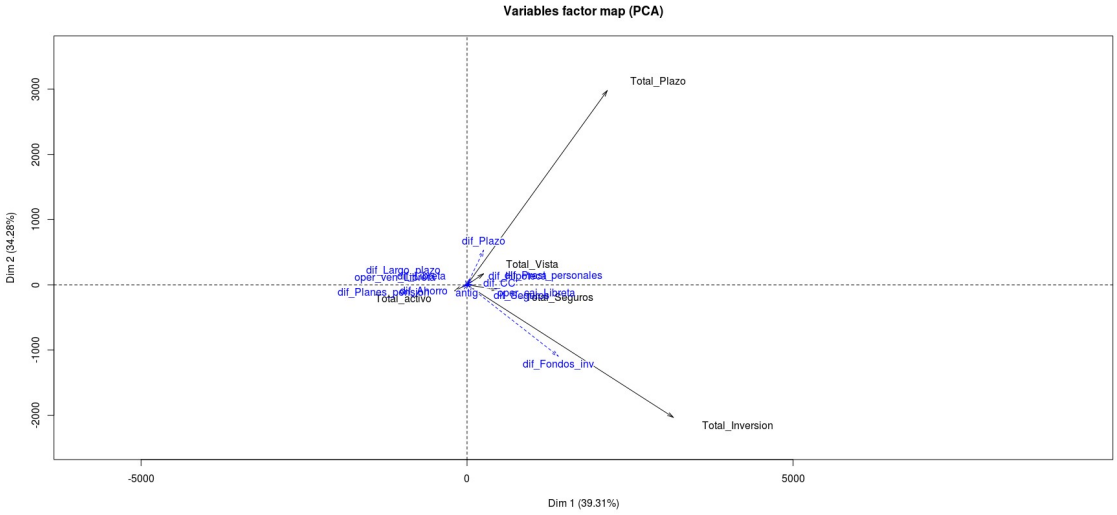
Para ello realizaremos primero un Análisis de Componentes Principales, para ver cuáles son los factores latentes que estructuran los datos y minimizar la parte de fluctuación aleatoria para a continuación efectuar el “Clustering” (en la 5ª sesión).

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.

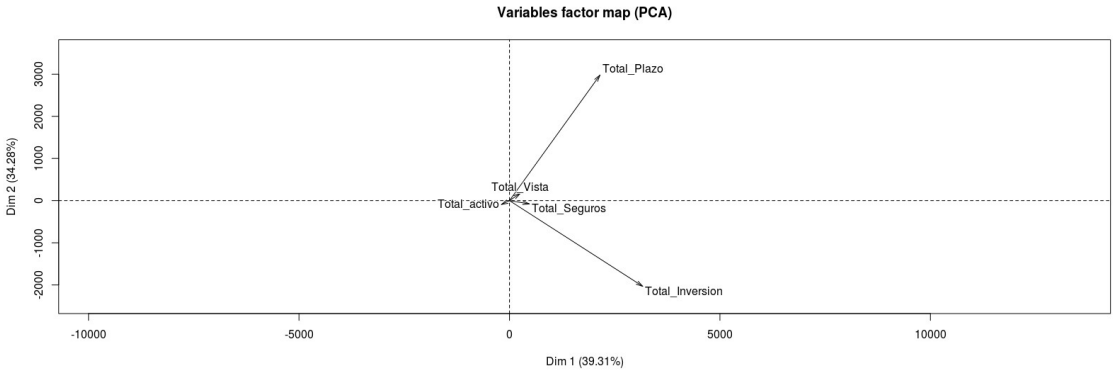
No vamos a estandarizar los datos, nos interesa que cada producto tenga su importancia en función de su valor.

```
 [1] “Baja”                  ← suplementaria cualitativa
 [2] "edatcat"               ← suplementaria cualitativa
 [3] "sexo"                  ← suplementaria cualitativa
 [4] "antig"                 ← suplementaria cuantitativa
 [5] "Nomina"                ← suplementaria cualitativa
 [6] "Pension"               ← suplementaria cualitativa
 [7] "Debito_normal"         ← suplementaria cualitativa
 [8] "Debito_aff"            ← suplementaria cualitativa
 [9] "VISA"                  ← suplementaria cualitativa
[10] "VISA_aff"              ← suplementaria cualitativa
[11] "MCard"                 ← suplementaria cualitativa
[12] "Amex"                  ← suplementaria cualitativa
[13] "Total_activo"          ←
[14] "Total_Plazo"           ←
[15] "Total_Inversion"       ←       ACTIVE
[16] "Total_Seguros"         ←
[17] "Total_Vista"           ←
[18] "dif_resid"             ← suplementaria cualitativa
[19] "oper_caj_Libreta"      ← suplementaria cuantitativa
[20] "oper_ven_Libreta"      ← suplementaria cuantitativa
[21] "dif_CC"                ← suplementaria cuantitativa
[22] "dif_Libreta"           ← suplementaria cuantitativa
[23] "dif_Plazo"             ← suplementaria cuantitativa
[24] "dif_Ahorro"            ← suplementaria cuantitativa
[25] "dif_Largo_plazo"       ← suplementaria cuantitativa
[26] "dif_Fondos_inv"        ← suplementaria cuantitativa
[27] "dif_Seguros"           ← suplementaria cuantitativa
[28] "dif_Planes_pension"    ← suplementaria cuantitativa
[29] "dif_Hipoteca"          ← suplementaria cuantitativa
[30] "dif_Prest_personales"  ← suplementaria cuantitativa
```

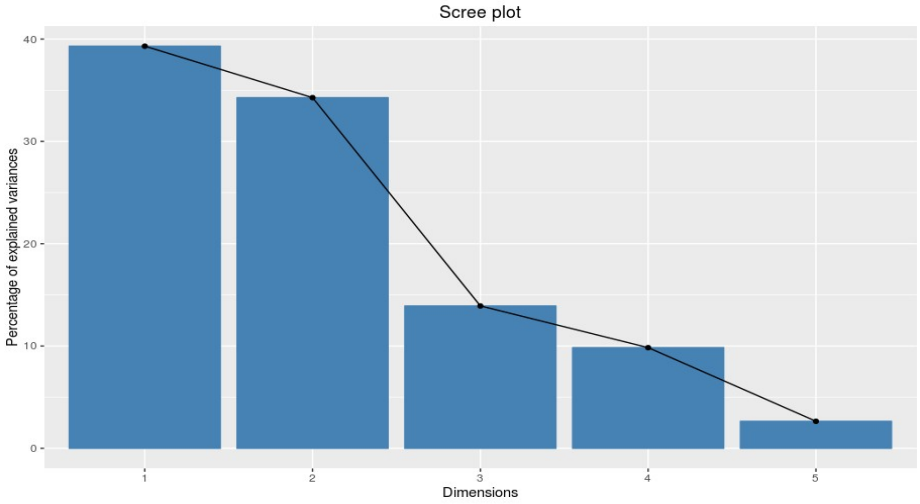
```
> pca.df <- PCA(df,ncp=10,quali.sup=c(1:3,5:12,18),quanti.sup=c(4,19:30), scale.unit = F)
```

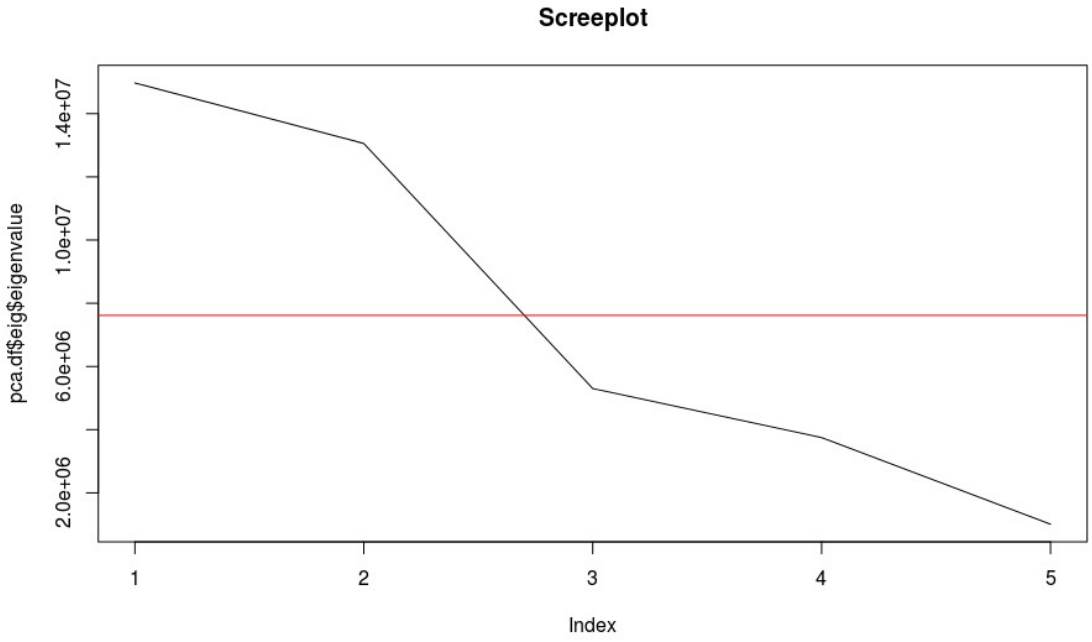


Detalle de variables activas:



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?.





Usando el 'last elbow rule' mediante el 'Screeplot', tenemos 2 dimensiones significativas.

Porcentaje de varianza acumulado:

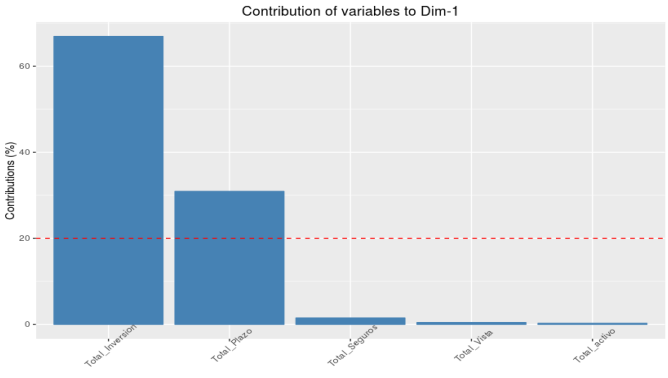
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	14970272	39.309051	39.30905
comp 2	13054677	34.279066	**73.58812**
comp 3	5300980	13.919352	87.50747
comp 4	3749254	9.844818	97.35229
comp 5	1008342	2.647712	100.00000

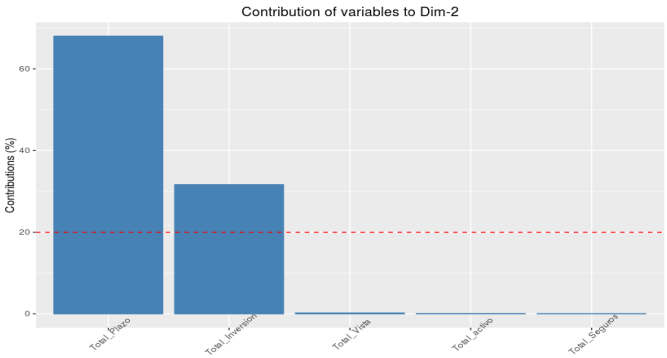
Habiendo tomado 2 dimensiones, el porcentaje de varianza acumulado es 73.59%.

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?.

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

Loadings:
                 Dim.1  Dim.2
Total_activo
Total_Plazo              0.999
Total_Inversion  0.996
Total_Seguros    0.217  0.109
Total_Vista              0.272
```

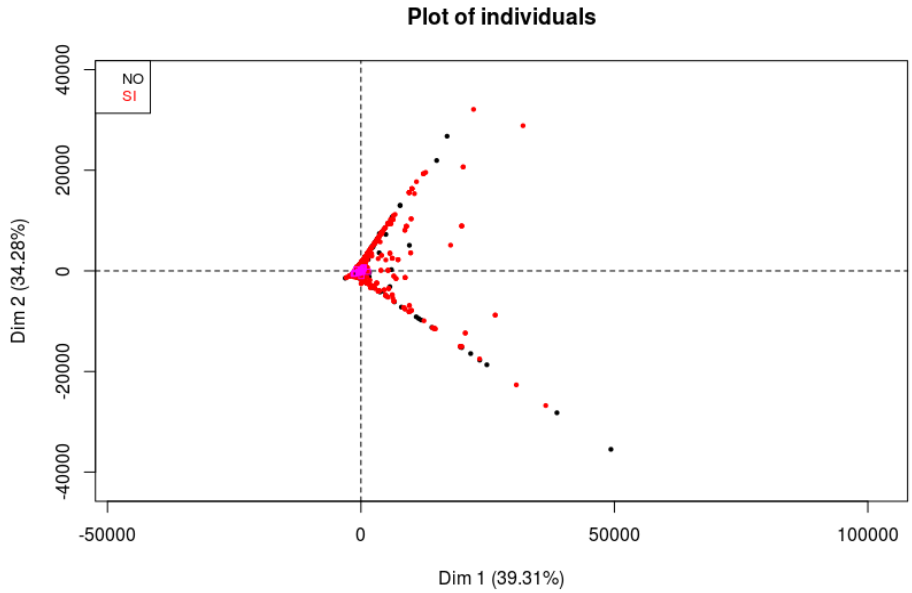




Factores latentes: Total_Plazo, Total_Inversion

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”?.

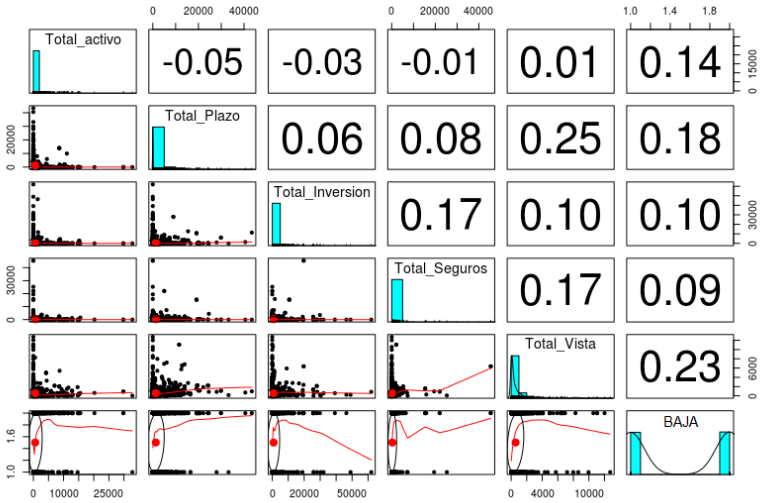
```
> plot(pca.df, axes = c(1, 2), choix = c("ind"), habillage=1, label="none", title="Plot of individuals",
cex=0.7)
```



No se observa ningún factor que claramente cause la “baja”.

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?

```
> df[18]=as.data.frame(as.numeric(df[[1]]))
> pairs.panels(df[13:18])
```



La variable "baja" tiene correlaciones, pero muy bajas, con las variables activas; considerando solamente las 2 primeras (Total_Plazo, Total_Inversion) su correlación con "baja" sigue siendo MUY baja (0.25, 0.17).