CLUSTERING PRACTICA

2024-04-15

Captura de datos

```
library(readr)
consumo_electrico <- read_csv("C:/Users/glady/Downloads/consumo_electrico.csv")
```

```
## Rows: 157 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (15): molienda_cereales_y_oleaginosas, resto_de_alimentos, bebidas, tab...
## date  (1): periodo
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

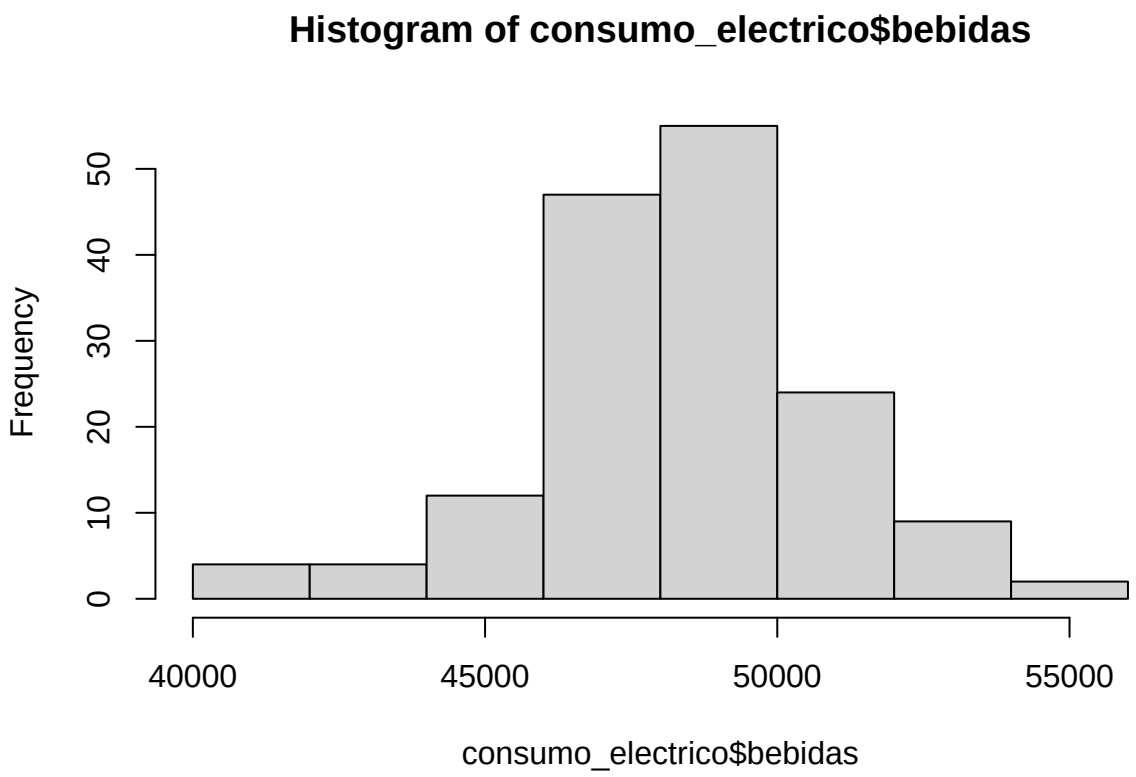
Como traer parte de la tabla

```
consumo_electrico[c(1:10,30:35),]
```

```
## # A tibble: 16 x 16
##    periodo    molienda_cereales_y_oleaginosas resto_de_alimentos bebidas tabaco
##    <date>                               <dbl>              <dbl>   <dbl>  <dbl>
##  1 2011-01-01                         116652.            136660.  47866.  2726.
##  2 2011-02-01                         113853.            137724.  48171.  2976.
##  3 2011-03-01                         121216.            136769.  45082.  2830.
##  4 2011-04-01                         116191.            139764.  47766.  3004.
##  5 2011-05-01                         112922.            137179.  46236.  2867.
##  6 2011-06-01                         106098.            136408.  46878.  3363.
##  7 2011-07-01                         107011.            136605.  47277.  2835.
##  8 2011-08-01                         111412.            137067.  47451.  2781.
##  9 2011-09-01                         119621.            139692.  46179.  2782.
## 10 2011-10-01                         121165.            140418.  48068.  2789.
## 11 2013-06-01                         108754.            144386.  49386.  2620.
## 12 2013-07-01                         104757.            144101.  48918.  2776.
## 13 2013-08-01                         106544.            143226.  47597.  2920.
## 14 2013-09-01                         105868.            143158.  48857.  2826.
## 15 2013-10-01                          96441.            143556.  49475.  2867.
## 16 2013-11-01                          92305.            143387.  49525.  2907.
## # i 11 more variables: textil_indumentaria_y_cuero <dbl>,
## #   madera_papel_y_edicion <dbl>, refinacion_de_petroleo <dbl>, quimicos <dbl>,
## #   caucho_y_plastico <dbl>, minerales_no_metalicos <dbl>,
## #   metales_basicos <dbl>, metalmecanica <dbl>, automotriz <dbl>,
## #   resto_de_industria <dbl>, total_industria <dbl>
```

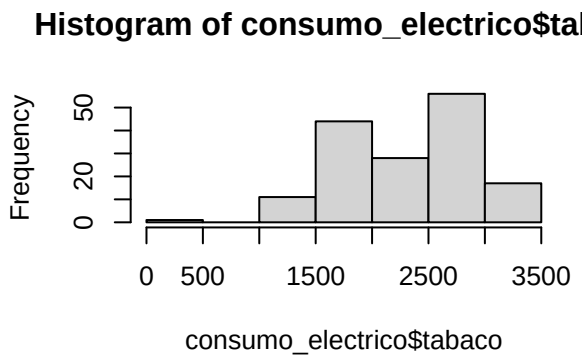
histograma de una Columna del dataset

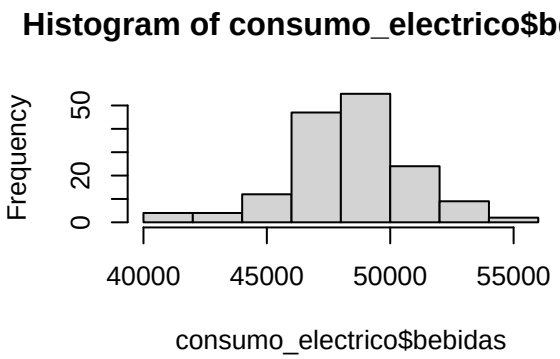
```
hist(consumo_electrico$bebidas)
```

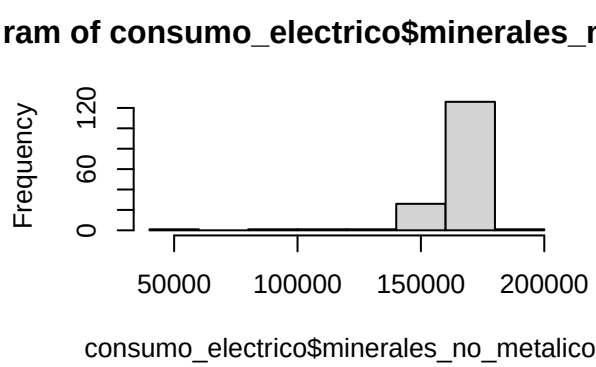


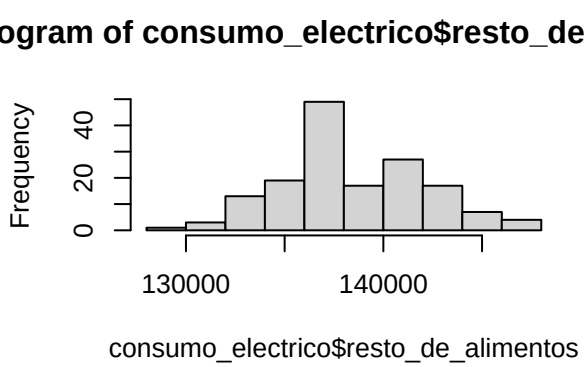
Histograma de 4 Colunnas a la vez, sirve para comparar

```
par(mfrow=c(2,2))
hist(consumo_electrico$tabaco)
hist(consumo_electrico$bebidas)
hist(consumo_electrico$minerales_no_metalicos)
hist(consumo_electrico$resto_de_alimentos)
```





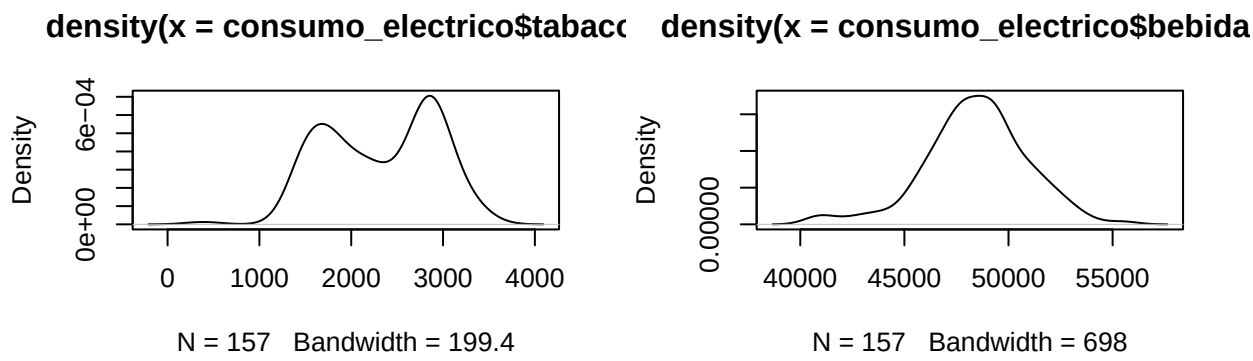


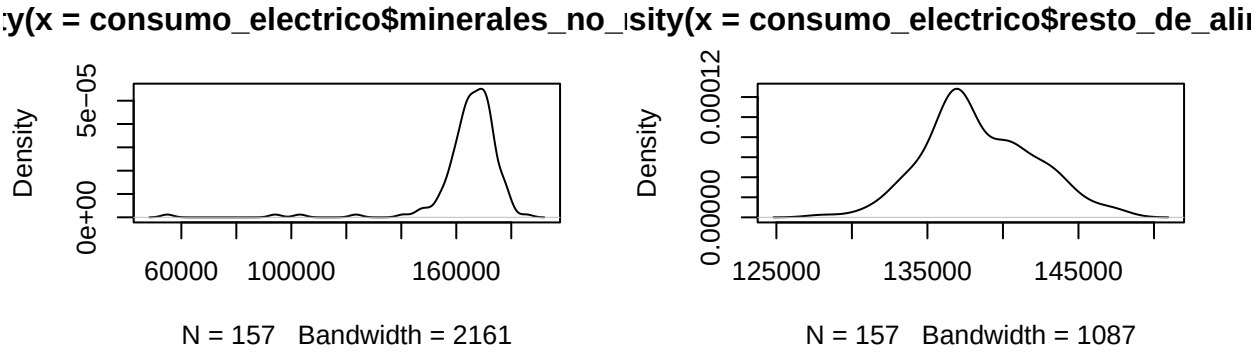


Gráficos de Densidad

Algunas personas prefieren utilizar la envolvente del histograma que es el gráfico de densidad

```
par(mfrow=c(2,2))
plot(density(consumo_electrico$tabaco))
plot(density(consumo_electrico$bebidas))
plot(density(consumo_electrico$minerales_no_metalicos))
plot(density(consumo_electrico$resto_de_alimentos))
```





Gráficas Ralas y Análisis Multivariado

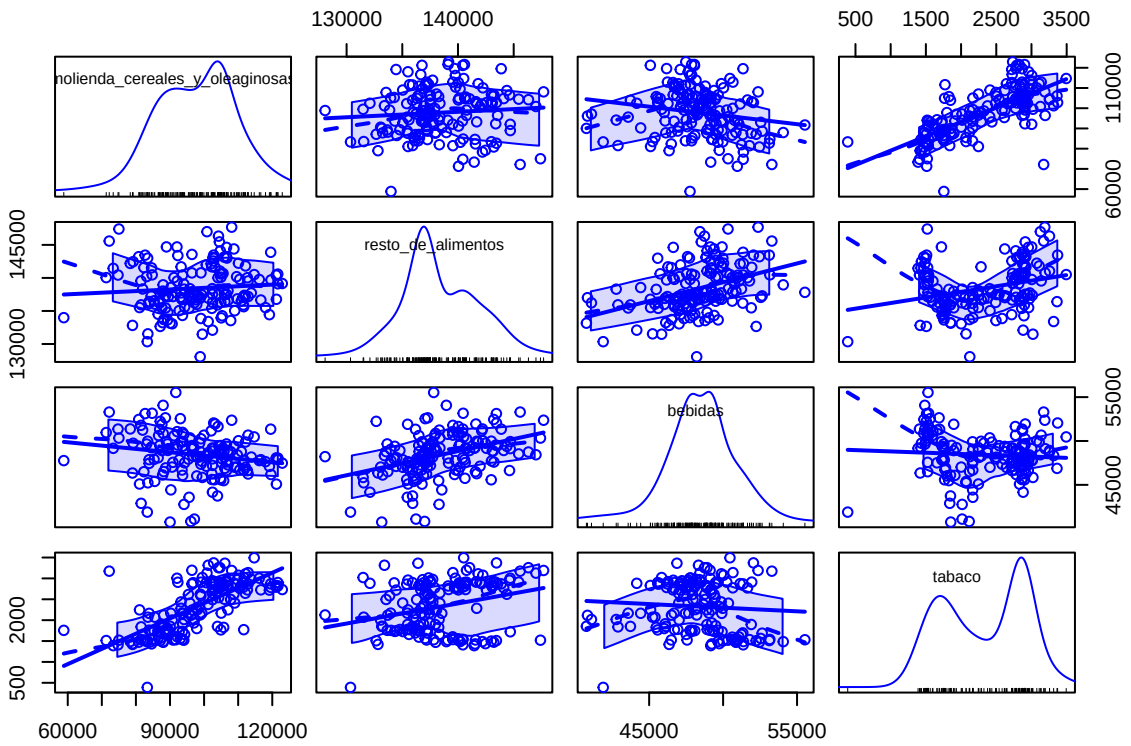
```
library(car)
```

```
## Loading required package: carData
```

```
consumoE_Rawdata <- consumo_electrico[ ,c(2,3,4,5)]
consumoE_Rawdata
```

```
## # A tibble: 157 x 4
##    molienda_cereales_y_oleaginosas resto_de_alimentos bebidas tabaco
##                              <dbl>              <dbl>   <dbl>  <dbl>
##  1                         116652.            136660.  47866.  2726.
##  2                         113853.            137724.  48171.  2976.
##  3                         121216.            136769.  45082.  2830.
##  4                         116191.            139764.  47766.  3004.
##  5                         112922.            137179.  46236.  2867.
##  6                         106098.            136408.  46878.  3363.
##  7                         107011.            136605.  47277.  2835.
##  8                         111412.            137067.  47451.  2781.
##  9                         119621.            139692.  46179.  2782.
## 10                         121165.            140418.  48068.  2789.
## # i 147 more rows
```

```
scatterplotMatrix(consumoE_Rawdata)
```



Mínimo numero de dimensiones

Cuándo nos enfrentamos a situaciones como esta, suele ocurrir que al definir los indicadores nos encontramos con el dilema del gran volumen de datos. Esto no es un problema que provenga tan solo del número de casos que estudiamos con el objeto de conocer el recorrido de una variable, sino más bien por la gran cantidad de variables o calificadores con los que los definimos o estudiamos. Ya vimos en el caso anterior como dimensiones o variables que tienen distinto nombre no son en realidad más que la misma cosa. En el ejemplo anterior la pregunta era si podríamos prescindir de una variable. En este ejercicio trataremos de ver cuantas podemos eliminar. La consigna es Mientras menos variables mejor, y la restricción que impondremos será la de perder variables siempre que podamos seguir describiendo con alto nivel de confianza el comportamiento de todos los casos. Otra mirada sobre el problema podría enunciarse así. “Como puedo saber que valores o recorrido le impondría a la mínima cantidad de variables para calificar como candidato interesante en la nómina de contratistas de las grandes empresas constructoras”.

Para auxilio en este problema utilizaremos el Método de Análisis de Componentes Principales. En este caso y al igual que en el caso anterior usaré un subconjunto de datos (sólo los numéricos) y en especial la matriz de correlación. Esta matriz está armada con las pendientes de las aproximaciones lineales de las rectas del gráfico de densidades.

Las técnicas que usaremos pretenden desde sus diferentes enfoques abrodar el problema de simplificar la interpretación del comportamiento individual y colectivo de los casos (empresas constructoras y contratista) y como podemos valernos del proceso de ingeniería inversa para mover los controles de nuestra “nave” en el tablero de comando con el que fijaremos la altura de la vara del tablero de control.

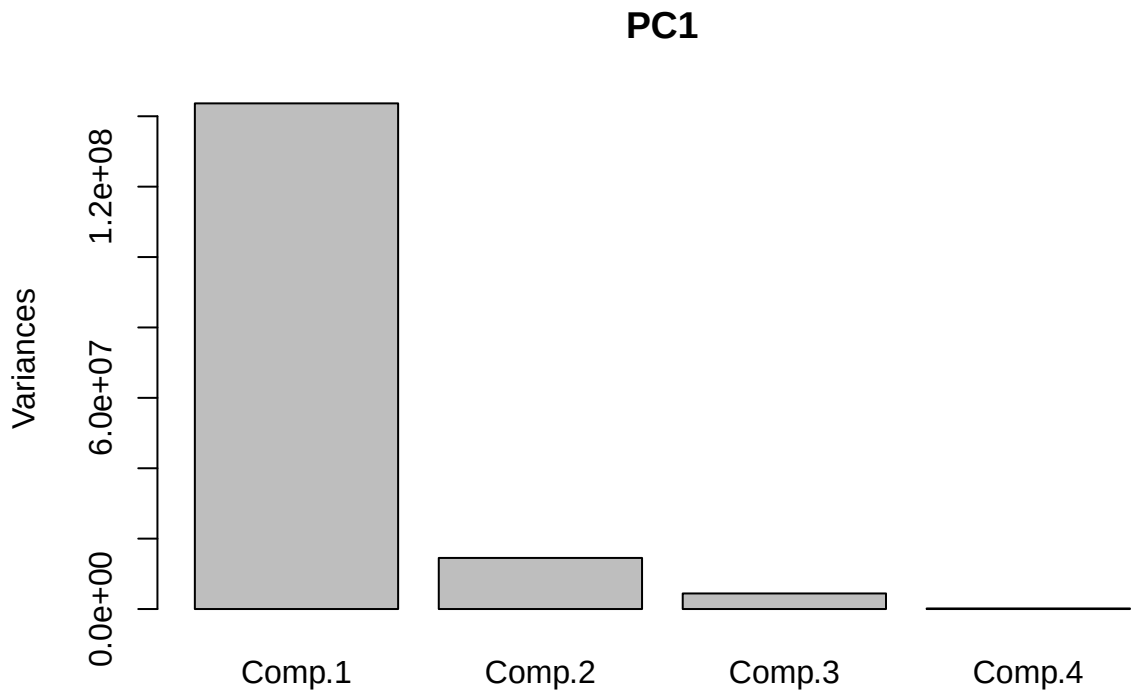
Análisis de Componentes Principales

Crearemos un objeto nuevo que se llamará PC1 (por Principal Component 1) y la instrucción con el que crearemos la matriz de correlaciones es princomp.

```
PC1 <- princomp(consumoE_Rawdata)
PC1
```

```
## Call:
## princomp(x = consumoE_Rawdata)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4
## 11985.751  3809.797  2103.983   397.294
##
##  4  variables and  157 observations.
```

```
plot(PC1)
```



En el ploteo podemos ver que uno de los componentes principales aporta casi el 4 veces más de la información referida al comportamiento de la varianza de todos los casos. Este componente es el que más incluye en la clasificación o posible identificación del comportamiento de cada individuo de la muestra.

«sumario_pc1,echo=TRUE»=

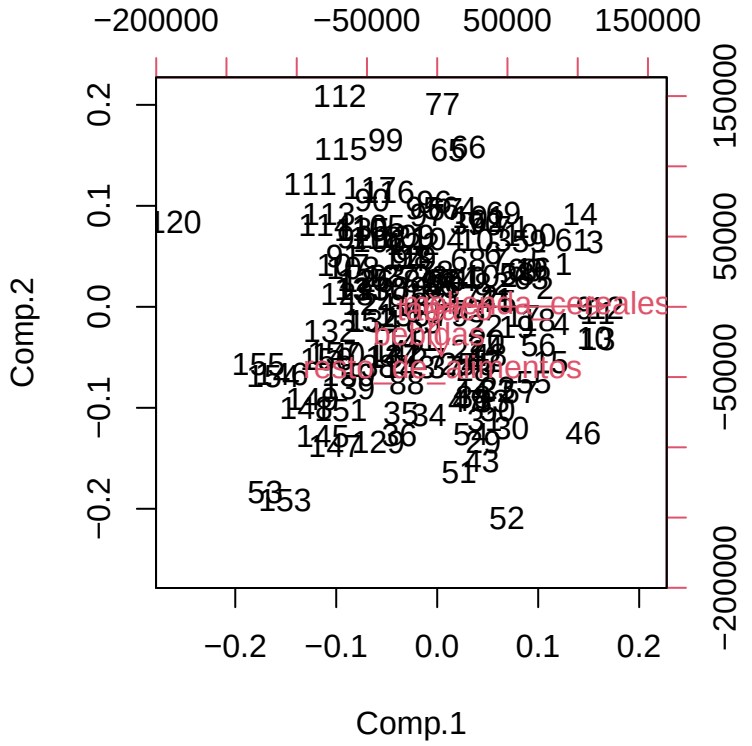
```
summary(PC1)
```

```
## Importance of components:
##                             Comp.1       Comp.2       Comp.3       Comp.4
## Standard deviation     1.198575e+04 3.809797e+03 2.103983e+03 3.972940e+02
## Proportion of Variance 8.826527e-01 8.917911e-02 2.719843e-02 9.698024e-04
## Cumulative Proportion  8.826527e-01 9.718318e-01 9.990302e-01 1.000000e+00
```

Si observamos bien el reporte que nos entrega el comando summary nos podemos dar cuenta que con los dos primeros componentes podríamos explicar 97.768521% del comportamiento de las muestras de la población. En nuestro caso del total de empresas contratistas analizadas.

¿Qué pasaría si representamos a las empresas en un gráfico en el que las variables de los ejes sean los dos componentes principales? , pues tendríamos un primer indicio de la bondad de las dimensiones o variables para agrupar a las muestras Esto lo podemos realizar con el comando bilot

```
biplot(PC1)
```



Los ńumeros que aparecen el el diagrama son el caso de estudio (renglón en que se encuentra la empresa contratista). A simple vista observamos que hay como tres tipos distintos empresas (tres nubes claramente diferenciadas). Aquí nos queda claro que el principal componente que ordena o divide a estas colonias es indistintamente el CAPITAL o el EQUIPAMIENTO con que cuentas.

También podemos ver que hay empresas como la 15, 16, 132, 118, 61, 107 sobre las que el gráfico nos recomienda estudiarlas más pues no es capaz de clasificarlas bien (son casos extremos o anómalos). Tal vez con poco capital o sin equipo pueden llegar a ser competitivas o interesantes para las grandes constructoras.

Por último la dimensión referida a la cetificación de NORMAS es la dimensión que menos valor aporta. Esto no implica que no certificar sea poco importante, sino que probablemente sea una pregunta irrelevante si todos contestaron que SI certificaron ISO 9000.

Scores

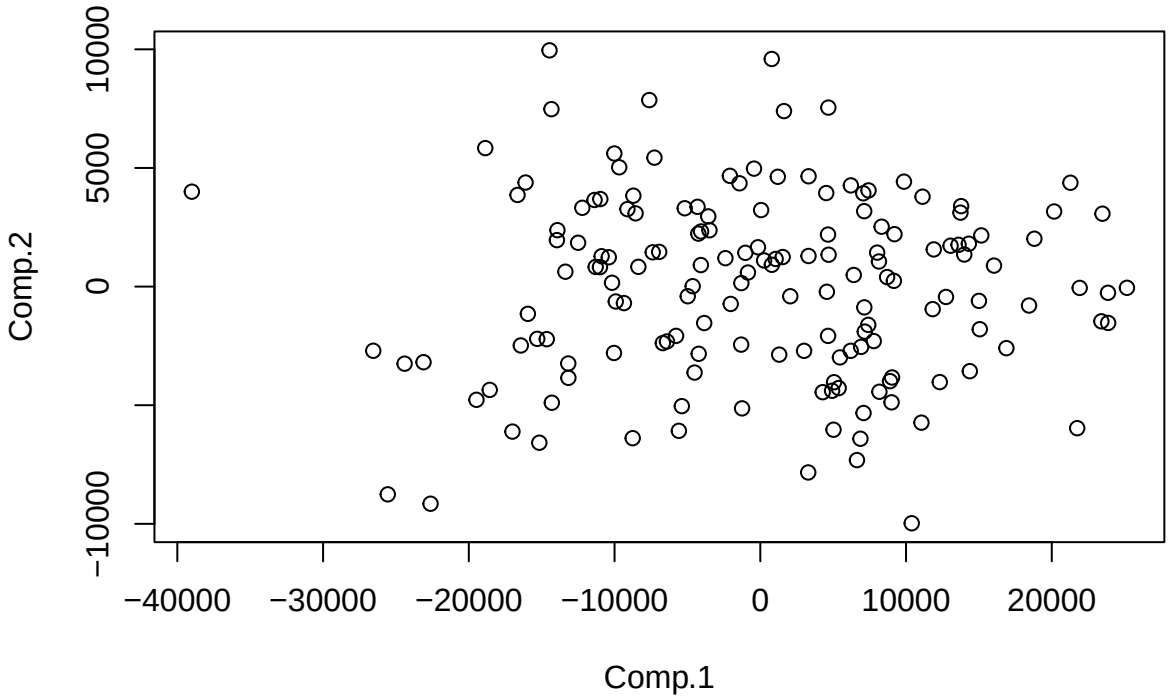
Si el comportamiento del componente va hacia el lado positivo, se debe interpretar como que a mayor desempeño mejor resultado o calificación. Si algún componente apunta para el lado negativo tendremos que pensar que a mayor calificación en esa dimensión pero sería el desempeño. La variable PC1 que usamos tiene mucha información valiosa. Revise todo el contenido, voy a mostrar una dimensión que es el score que indica como se comportarían todos los individuos si sólo los analizásemos con los componentes 1 y 2.

```
acp1 <- PC1$scores
acp1 [1:10 , ]
```

```
##          Comp.1      Comp.2      Comp.3     Comp.4
##  [1,] 18804.697  2016.27206  1079.13060  190.56431
##  [2,] 16035.222   884.82205   794.72907 -115.95905
##  [3,] 23475.237  3072.28262 -1300.98384  282.09978
##  [4,] 18438.721  -798.33257  -282.72273   28.67101
##  [5,] 15163.097  2155.54480  -798.15532  -40.70699
##  [6,]  8324.273  2523.52087  -235.02264 -814.30046
##  [7,]  9206.453  2208.88478   107.96924 -251.09662
##  [8,] 13602.603  1759.73216   289.92892  -25.82768
##  [9,] 21914.972   -56.48778 -1539.47164  385.90191
## [10,] 23401.703 -1464.47950   -29.00679  441.85253
```

Voy a realizar el mismo score pero ahora solo con los componentes 1 y 2

```
acp2 <-PC1$scores[ ,1:2]
plot(acp2)
```



Aquí ya podemos ver más claramente la división que se produce entre distintos clusters. Para poder diferencias aún más recurriremos a un nuevo tipo de análisis diferenciado que se llama análisis de clusters

Análisis de Clusters o Conglomerados

Para realizar este análisis recurriremos a cargar la biblioteca clusters

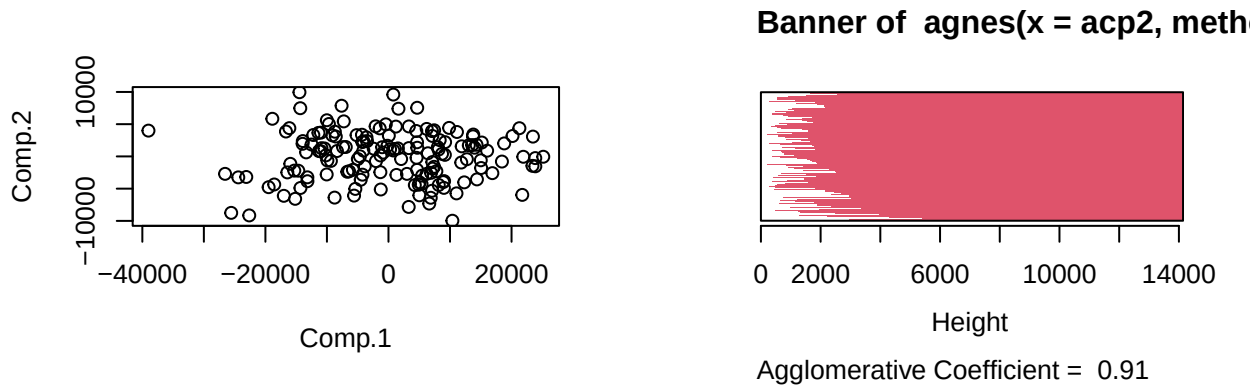
En el análisis de conglomerados existen dos formas clásicas de estudio. Ambas recurren a las distancias euclídeas entre las muestras. Tenemos aproximaciones Jerárquicas y No Jerárquicas AGNES, CLARA, DIANA, MORA, PAM son nombres de las técnicas que la biblioteca Clusters usa. Todas las técnicas se caracterizan por ser un acrónimo de la combinación de aproximaciones que usan (Single Linkage, Complete Linkage, Average Linkage) .

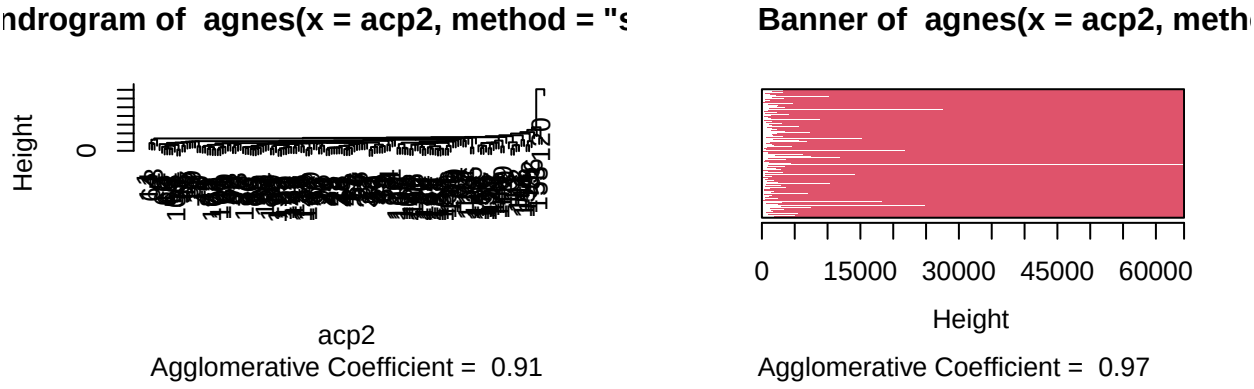
Todas tienen nombre de mujer, pero esto no quiere necesariamente decir que se trate de una técnica con complicaciones inesperadas, sino más bien que si quieres lo mejor de una de ellas es mejor que la entiendas e indagues en la página del manual.

```
library(cluster)
agp1 = agnes(acp2,method="single")
agp2 = agnes(acp2,method="complete")
agp3 = agnes(acp2,method="average")
```

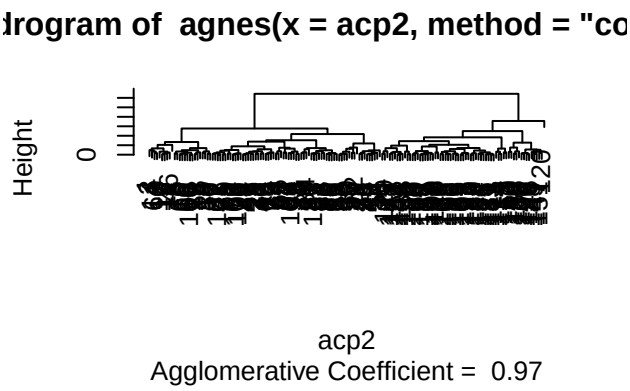
Con la clasificacion terminada procederemos a ver gráficamente el resultado.

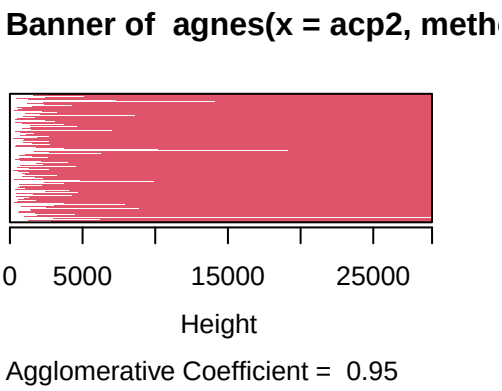
```
par(mfrow=c(2,2))
plot(acp2)
plot(agp1)
plot(agp2)
```

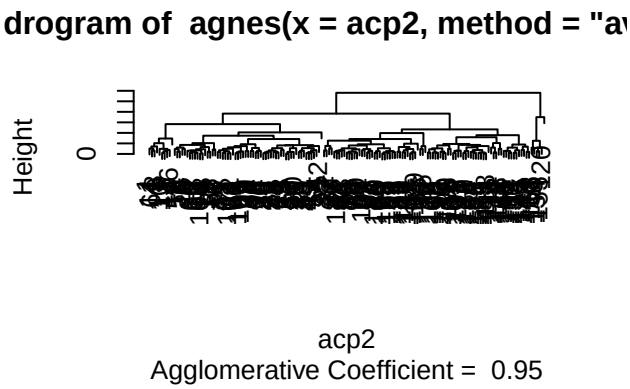





```
plot(agp3)
```

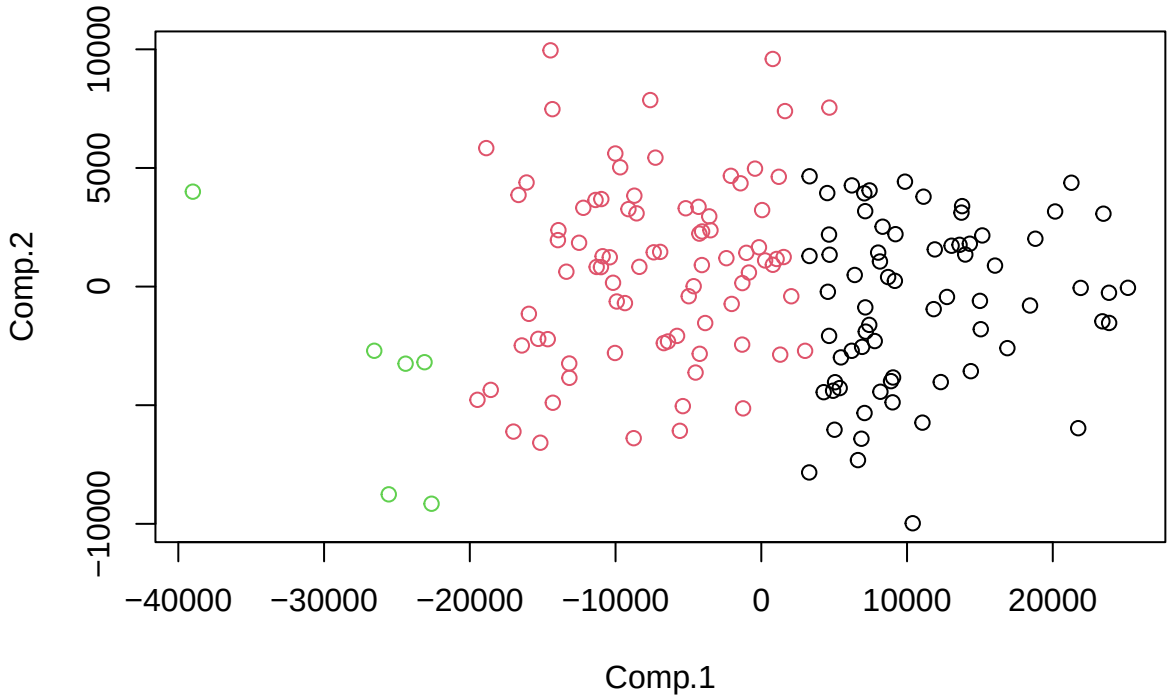






Pasa asignar las muestras a grupos usaré el comendo cuttree que me permite valerme de las franjas blancas de corte del los gráficos para armar los clusters

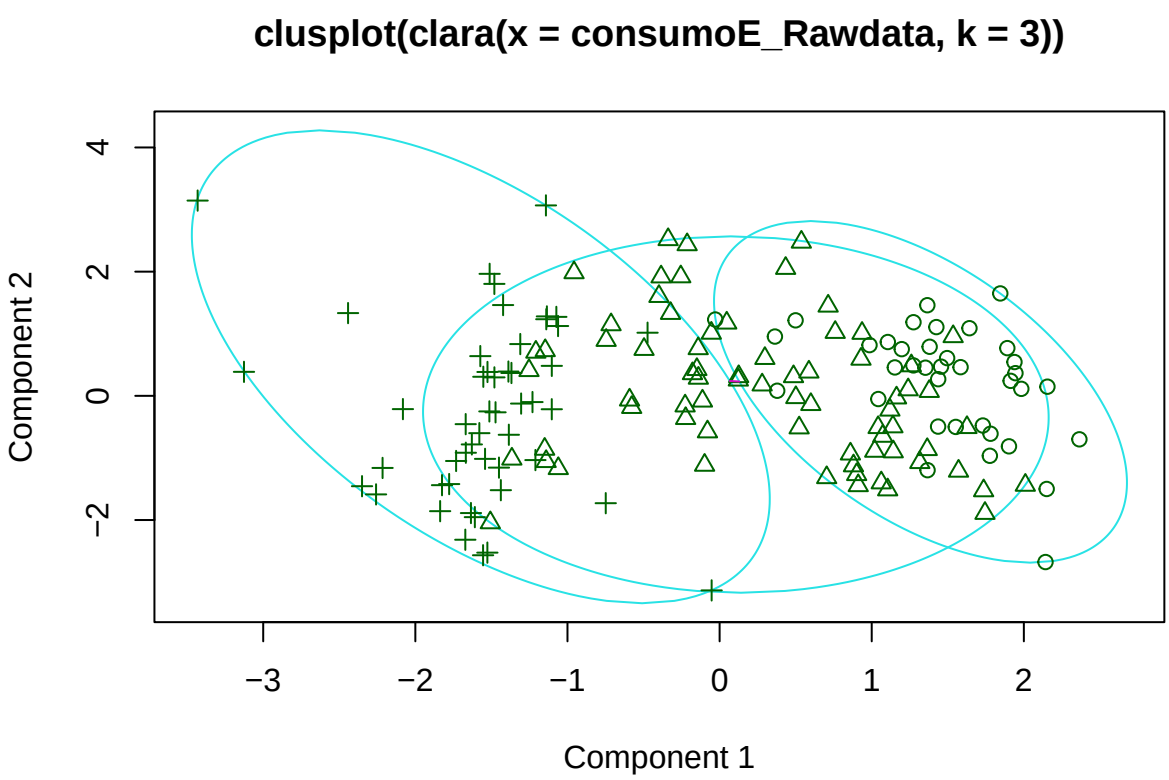
```
agpcut <- cutree(agp3,3)
par(mfrow=c(1,1))
plot(acp2,col=agpcut)
```

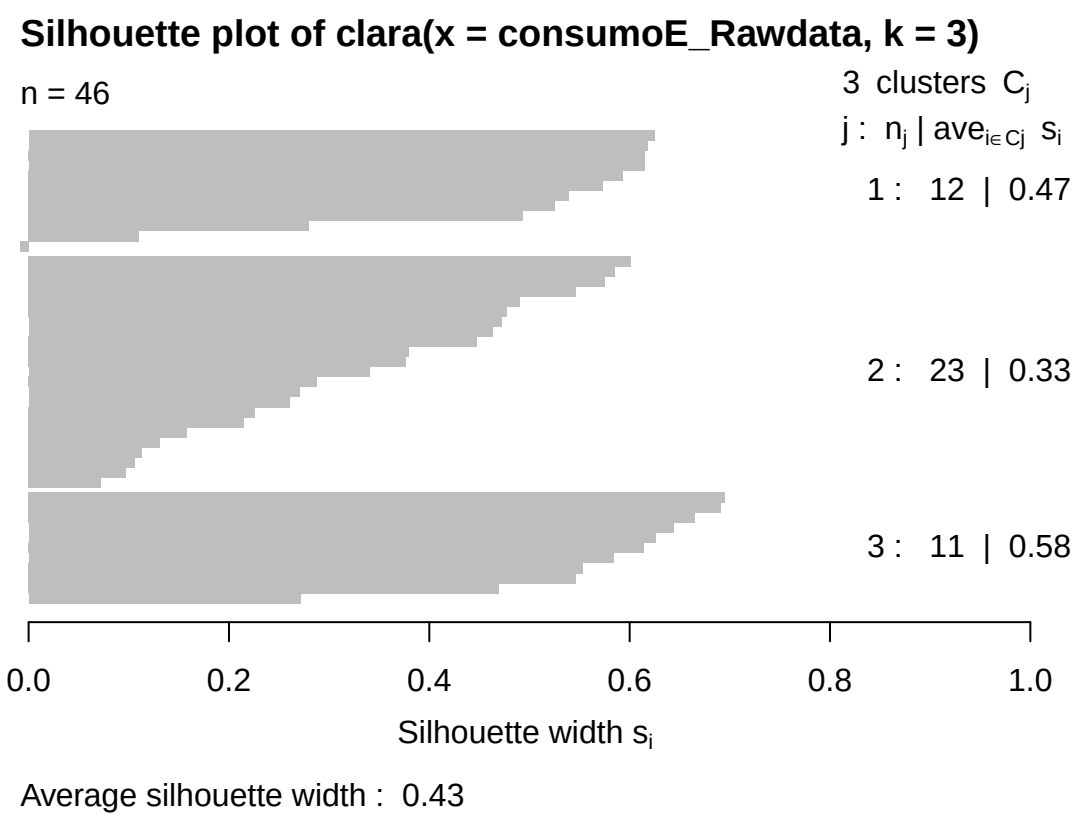


Otros gráficos de agrupamiento

Método Clara

```
plot(clara(consumoE_Rawdata,3))
```

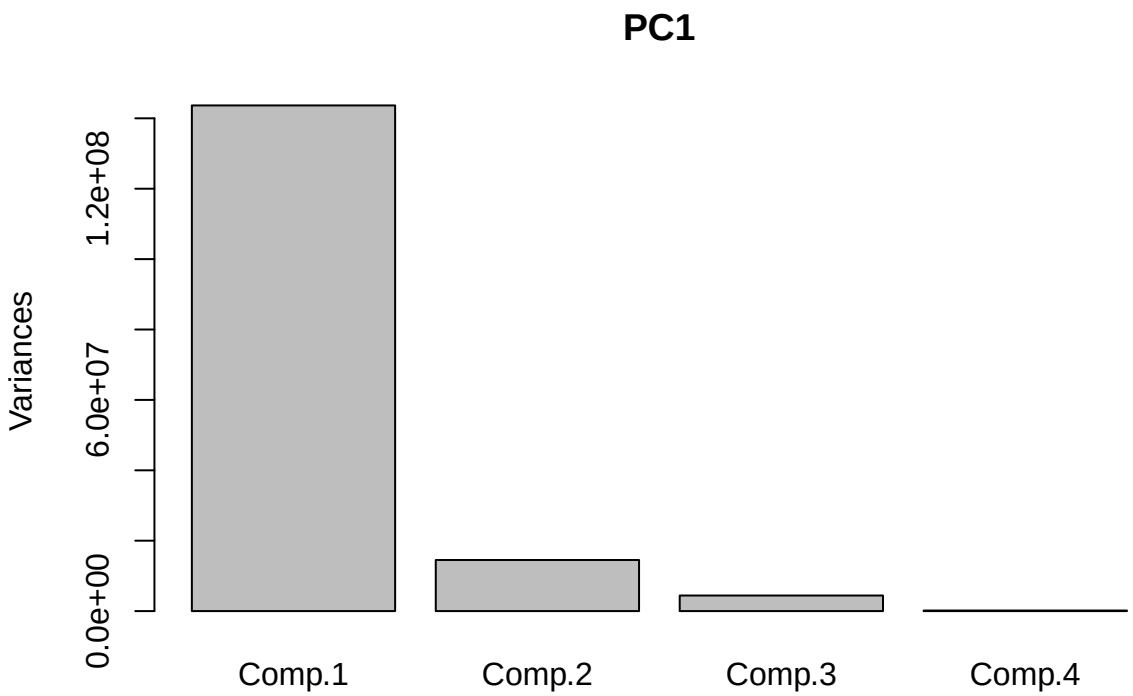

clusplot(clara(x = consumoE_Rawdata, k = 3))
Component 2
4
2
0
−2
−3
−2
−1
0
1
2
Component 1
These two components explain 80.43 % of the point variability.



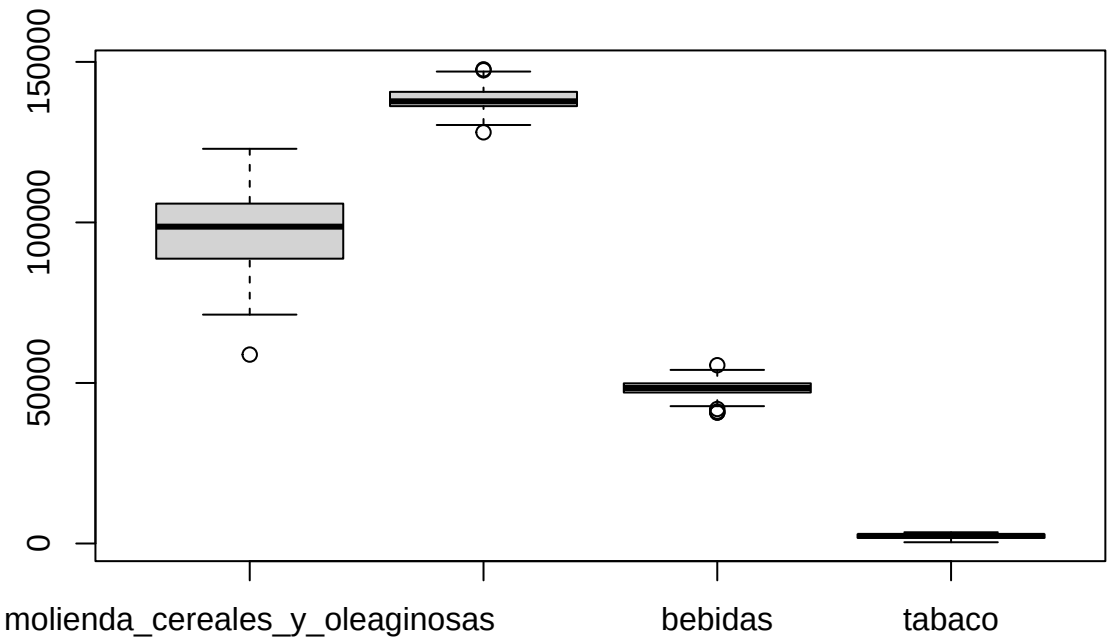

```
PC1 <- princomp(consumoE_Rawdata)
PC1
```

```
## Call:
## princomp(x = consumoE_Rawdata)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4
## 11985.751  3809.797  2103.983   397.294
##
##  4  variables and  157 observations.
```

```
plot(PC1)
```




```
boxplot(consumoE_Rawdata)
```


150000
100000
50000
0
molienda_cereales_y_oleaginosas
bebidas
tabaco