Técnicas de análisis Multivariado

Alumnos:
Huertas Quispe, Anthony Enrique ***Cod:*** *20173728*
Torres Salinas, Karina Hesi ***Cod:*** *20164111*
Córdova Proleón, Christian Therius ***Cod:*** *20173970*

Semestre: *2017-II*

Tema: Lista 1

Prof. Enver Tarazona

Pontificia Universidad Católica del Perú
Escuela de Posgrado
Maestría en Estadística

Ejercicio 1 (7 puntos)

En el archivo de SPSS DepartamentosPeru.sav se presenta información relacionada a las siguientes variables de los departamentos del Perú:

- Departamento: Nombre del departamento
- Vida: Esperanza de vida al nacer (años)
- Alfabetismo: Tasa de alfabetismo de adultos (%)
- Ingreso: Ingreso familiar per cápita (S/. Mes)
- Identidad: % Población con acta de nacimiento o DNI
- Salud: Tasa de escolaridad de 5 a 18 años (%)
- Saneamiento: % Viviendas con acceso a agua y desagüe a la vez
- Electrificación: % Viviendas con electricidad
- Policía: Policías por cada mil habitantes

a) Usando el método de K-medias y PAM, determine el número adecuado de conglomerados que se deberían usar para agrupar a los departamentos. Utilice al menos cinco criterios de comparación que sustenten su respuesta. (**3 puntos**)

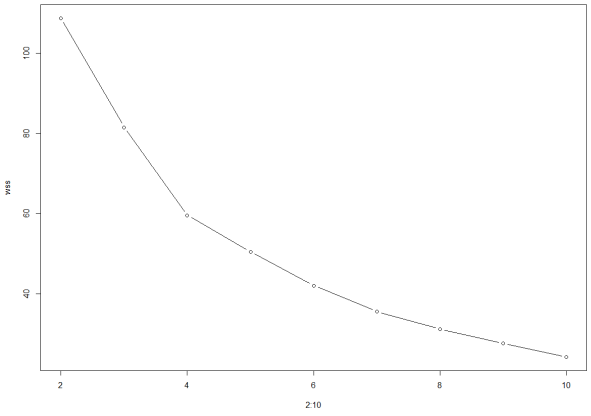
Nota: Código **ANEXO 1**.

La base departamentos presenta algunas variables que están medidas en escalas distintas, por tal se hará una transformación de las variables para homogeneizar las escalas, esto con el fin de que no afecte la conformación final de los conglomerados; a continuación presentamos los resultados de los criterios de comparación.

Listing 1: **Criterio 1**. Suma de cuadrados dentro de los clúster.

```
wss<-numeric()
for(h in 2:10){
b<-kmeans(scale(departamentos),
          h,nstart = 20)
wss[h-1]<-b$tot.withinss
}
plot(2:10,wss,type="b")
wss
```

108.70448	81.50711	59.56630	50.48723
42.02720	35.56708	31.16218	27.73415
24.24449			

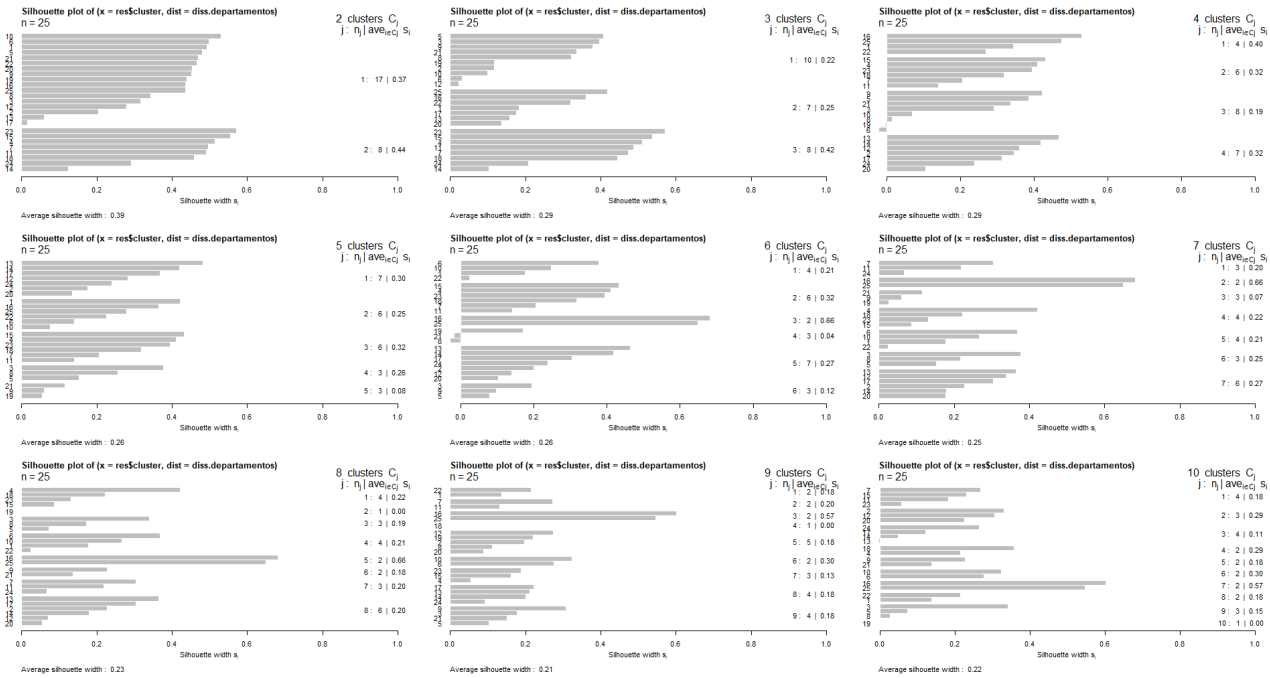


Del gráfico de sedimentación podemos sugerir usar 3 ó 4 conglomerados para agrupar a los departamentos, a partir del conglomerado 4 la ganancia en términos de varianza se hace más pequeña.

Listing 2: **Criterio 2**. Silueta.

```
kmeansruns(scale(departamentos),criterion="asw")
```

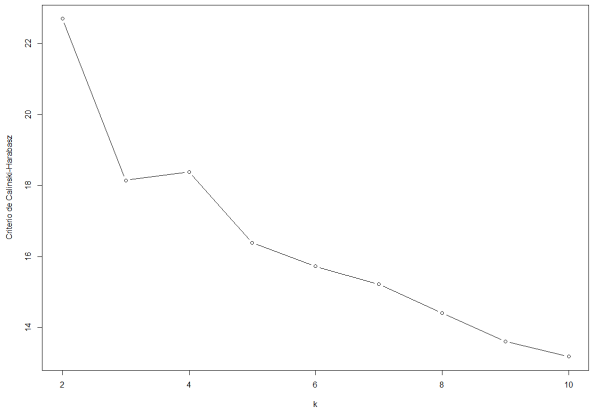
```
K-means clustering with 2 clusters of sizes 17, 8

Cluster means:
        vida alfabetismo escolaridad    ingreso  identidad      salud saneamiento
1 -0.5603059  -0.4306735  -0.4150396 -0.4984779 -0.3405792 -0.5063146  -0.5310237
2  1.1906500   0.9151812   0.8819592  1.0592655  0.7237308  1.0759186   1.1284254
  electrificacion    policia
1      -0.5179855 -0.4205665
2       1.1007192  0.8937038

Clustering vector:
AMAZONAS      ANCASH       APURÍMAC      AREQUIPA      AYACUCHO      CAJAMARCA
           1           1           1           2           1           1
CALLAO        CUSCO       HUANCAVELICA  HUÁNUCO       ICA           JUNÍN
           2           1           1           1           2           1
LA LIBERTAD   LAMBAYEQUE    LIMA          LORETO        MADRE DE DIOS MOQUEGUA
           1           2           2           1           1           2
PASCO         PIURA         PUNO          SAN MARTÍN    TACNA         TUMBES
           1           1           1           1           2           2
UCAYALI
           1

Within cluster sum of squares by cluster:
[1] 80.85455 27.84992
 (between_SS / total_SS =  49.7 %)

Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
 [6] "betweenss"    "size"         "iter"         "ifault"       "crit"
[11] "bestk"
```



De la gráfica anterior podemos observar la solución de 2 a 10 conglomerados, el promedio del índice de silueta para la primera solución con 2 conglomerados es de 0.39, siendo el mayor índice de todos, se puede tomar esto como una buena solución.

Listing 3: **Criterio 3**. Calinski-Harabasz.

```
kmeansruns(scale(departamentos),
           criterion="ch")
```




```
K-means clustering with 2 clusters of sizes 17, 8

Cluster means:
        vida alfabetismo escolaridad    ingreso  identidad      salud saneamiento
1 -0.5603059  -0.4306735  -0.4150396 -0.4984779 -0.3405792 -0.5063146  -0.5310237
2  1.1906500   0.9151812   0.8819592  1.0592655  0.7237308  1.0759186   1.1284254
  electrificacion    policia
1      -0.5179855 -0.4205665
2       1.1007192  0.8937038

Clustering vector:
AMAZONAS      ANCASH        APURÍMAC      AREQUIPA      AYACUCHO      CAJAMARCA
           1           1           1           2           1           1
CALLAO        CUSCO         HUANCAVELICA  HUÁNUCO       ICA           JUNÍN
           2           1           1           1           2           1
LA LIBERTAD   LAMBAYEQUE    LIMA          LORETO        MADRE DE DIOS MOQUEGUA
           1           2           2           1           1           2
PASCO         PIURA         PUNO          SAN MARTÍN    TACNA         TUMBES
           1           1           1           1           2           2
UCAYALI
           1

Within cluster sum of squares by cluster:
[1] 80.85455 27.84992
 (between_SS / total_SS =  49.7 %)

Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
 [6] "betweenss"    "size"         "iter"         "ifault"       "crit"
[11] "bestk"
```

En el índice de Calinski, la idea es que la suma de cuadrados entre los grupos sea mayor a la suma de cuadrados dentro de los grupos, a mayor valor de este indicador significa que hay una mejor conformación de los conglomerados, a menor valor significa que la conformación no es muy buena. De la salida observamos que índice de Calinski, alcanza el mayor valor con 2 conglomerados.

Listing 4: **Criterio 4 y 5**. Medidas de Validación Interna.

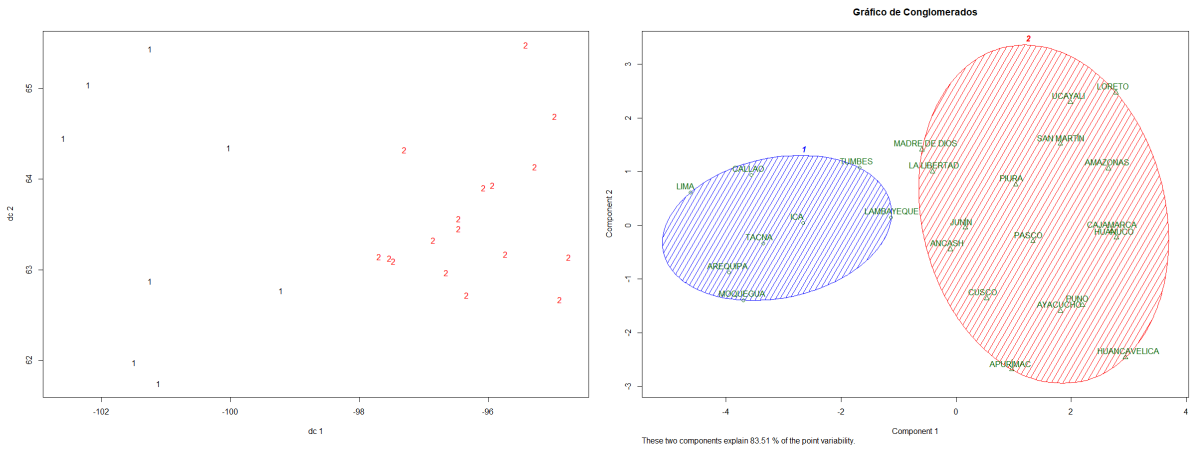
```
summary(intern)
```

```
Clustering Methods:
 kmeans

Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                            2       3       4       5       6       7       8       9      10

kmeans Connectivity    5.2333 12.2333 11.9667 16.1333 18.7000 21.3333 22.9000 26.4333 30.7833
       Dunn            0.2682  0.3415  0.4072  0.4072  0.4815  0.4807  0.5156  0.4806  0.5905
       Silhouette      0.3949  0.2768  0.2962  0.2738  0.2852  0.2515  0.2239  0.2081  0.1994

Optimal Scores:

             Score  Method Clusters
Connectivity 5.2333 kmeans 2
Dunn         0.5905 kmeans 10
Silhouette   0.3949 kmeans 2
```

La conectividad indica el grado de conexión de los conglomerados, como viene determinado por los k vecinos más cercanos (en este caso 5). Tiene un valor entre 0 e infinito y debe ser minimizado. De acuerdo a este criterio podríamos usar 2 conglomerados.
El índice de Dunn identifica un conjunto de clústeres que sean compactos, con una varianza pequeña entre las observaciones de los clústeres y que estos estén bien separados de las observaciones de otros clústeres. Sus valores van desde 0 hasta infinito y debe ser maximizado. Según este criterio podríamos usar 10 clústers.
De los gráficos podemos observar que la mayoría de los criterios coinciden en que el número adecuado de conglomerados es 2, se tiene la siguiente conformación:

Listing 5: Gráficos.

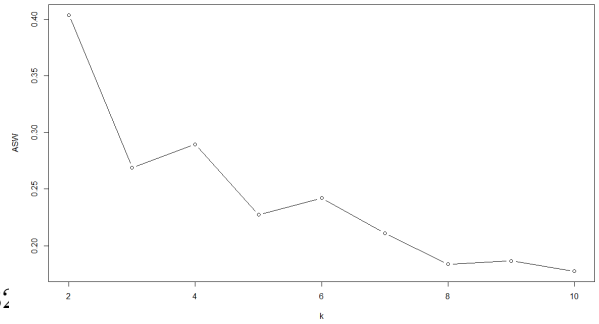
```
res=kmeans(scale(departamentos),2, nstart = 20)
plotcluster(departamentos,res$cluster)

clusplot(departamentos,res$cluster, color = TRUE,shade = TRUE, labels =2,lines=0,
         main ="Grafico de Conglomerados")
```



Método PAM:
Este método se basa en buscar k observaciones representativas entre todas las observaciones de un conjunto de datos, estas observaciones son llamadas medoides y una vez encontrados se construyen los k conglomerados asignando cada observación al medoide más cercano.

Listing 6: **Criterio 1 (PAM)**. Suma de cuadrados dentro de los clúster.

```
asw<-numeric()
for(h in 2:10){
res<-pam(scale(departamentos),h)
asw[h-1]<-res$silinfo$avg.width
}
plot(2:10,asw,type="b",xlab="k",ylab="ASW
    ")
asw
```




```
0.4038004 0.2689572 0.2895650 0.227453
0.2419963 0.2110292 0.1836160 0.186625
0.1774217
```

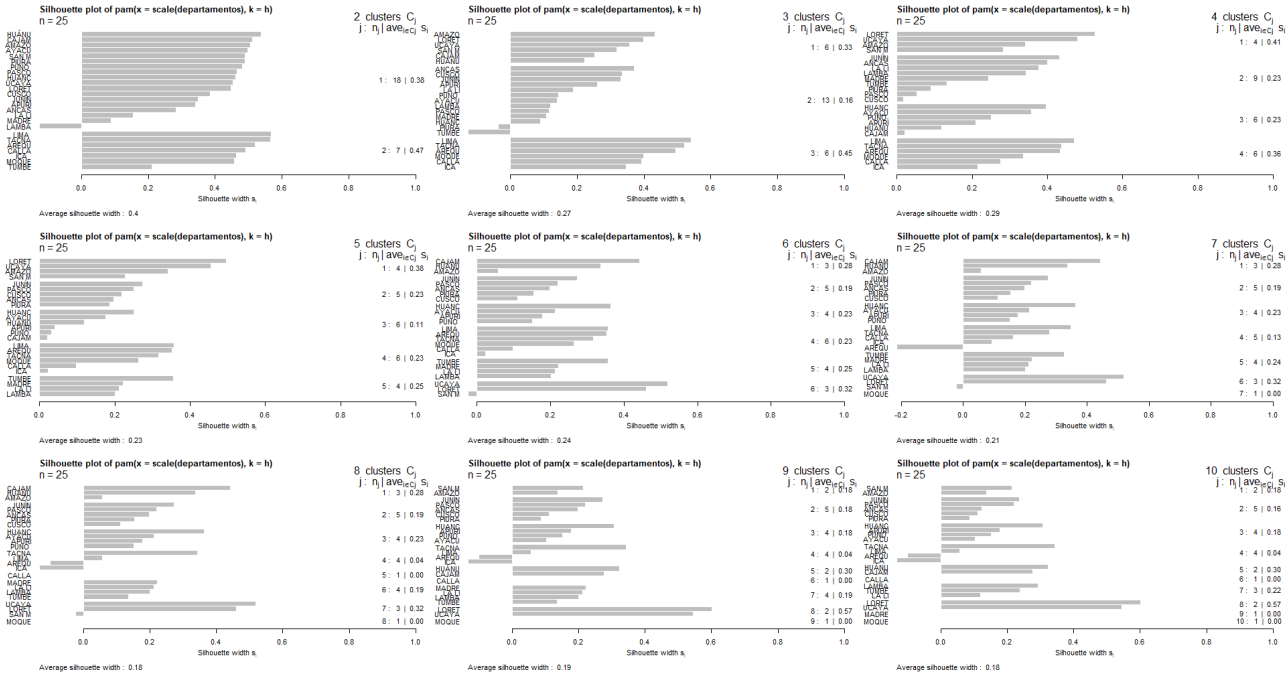
Aplicamos nuevamente el criterio de silueta, el mayor índice se obtiene con 2 conglomerados, el medoide en el primer conglomerado es Piura y en el segundo conglomerado es Tacna.

Listing 7: **Criterio 2 (PAM)**. Silueta.

```
pamk(scale(departamentos),criterion="asw")
```

```
$pamobject
Medoids:
              ID       vida alfabetismo escolaridad     ingreso  identidad      salud
PIURA         20 -0.4191064 -0.02473144  -0.6950451 -0.01115714 -0.3477906 -0.6442266
TACNA         23  0.8013459  0.90895817   1.0994498  0.90185480  0.9868412  1.7868813
              saneamiento electrificacion    policia
PIURA         -0.09855423     -0.08234677 -0.9018772
TACNA          1.45574300      0.98423092  0.8548219
Clustering vector:
AMAZONAS      ANCASH        APURÍMAC      AREQUIPA      AYACUCHO      CAJAMARCA
           1           1           1           2           1           1
CALLAO        CUSCO         HUANCAVELICA  HUÁNUCO       ICA           JUNÍN
           2           1           1           1           2           1
LA LIBERTAD   LAMBAYEQUE    LIMA          LORETO        MADRE DE DIOS MOQUEGUA
           1           1           2           1           1           2
PASCO         PIURA         PUNO          SAN MARTÍN    TACNA         TUMBES
           1           1           1           1           2           2
UCAYALI
           1
Objective function:
   build     swap
2.250457 2.154353

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"  "clusinfo"
 [7] "silinfo"    "diss"       "call"       "data"
```

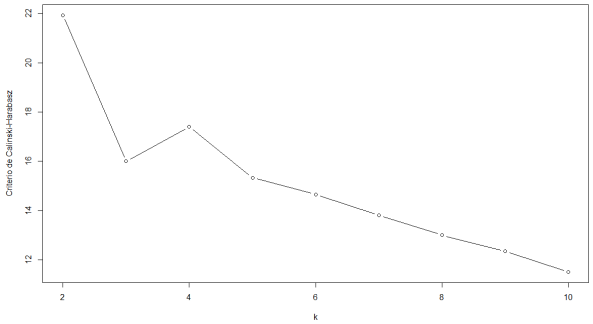
```
$nc
[1] 2

$crit
 [1] 0.0000000 0.4038004 0.2689572 0.2895650 0.2274532 0.2419963 0.2110292 0.1836160
 [9] 0.1866259 0.1774217
```



De la gráfica anterior, podemos observar la solución de 2 a 10 conglomerados, el promedio del índice de silueta para la primera solución con 2 conglomerados es de 0.4, siendo el mayor índice de todos.

Listing 8: **Criterio 3 (PAM)**. Calinski-Harabasz.

```
pamk(scale(departamentos),criterion="ch")
```




```
$pamobject
Medoids:
             ID       vida alfabetismo escolaridad     ingreso  identidad      salud
PIURA        20 -0.4191064 -0.02473144  -0.6950451 -0.01115714 -0.3477906 -0.6442266
TACNA        23  0.8013459  0.90895817   1.0994498  0.90185480  0.9868412  1.7868813
             saneamiento electrificacion    policia
PIURA         -0.09855423     -0.08234677 -0.9018772
TACNA          1.45574300      0.98423092  0.8548219
Clustering vector:
AMAZONAS      ANCASH        APURÍMAC      AREQUIPA      AYACUCHO      CAJAMARCA
           1           1           1           2           1           1
CALLAO        CUSCO         HUANCAVELICA  HUÁNUCO       ICA           JUNÍN
           2           1           1           1           2           1
LA LIBERTAD   LAMBAYEQUE    LIMA          LORETO        MADRE DE DIOS MOQUEGUA
           1           1           2           1           1           2
PASCO         PIURA         PUNO          SAN MARTÍN    TACNA         TUMBES
           1           1           1           1           2           2
UCAYALI
           1
Objective function:
   build     swap
2.250457 2.154353

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"  "clusinfo"
 [7] "silinfo"    "diss"       "call"       "data"

$nc
[1] 2

$crit
 [1]  0.00000 21.93597 16.00166 17.40103 15.34021 14.65467 13.81540 13.00384 12.34546
[10] 11.50281
```

De la salida observamos que el mayor índice Calinski se obtiene con 2 conglomerados, cuyos medoides son Piura para el primero y Tacna para el segundo (la solución es igual a la de criterio de siluetas).

Listing 9: **Criterio 4 y 5**. Medidas de Validación Interna.

```
summary(intern)
```

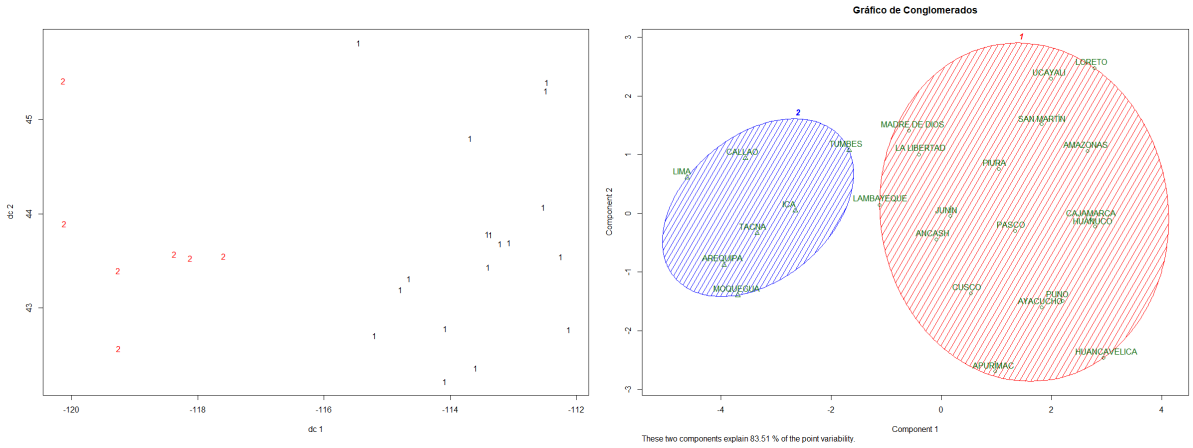
```
Optimal Scores:

             Score  Method Clusters
Connectivity 3.8667 pam    2
Dunn         0.4954 pam    8
Silhouette   0.4038 pam
```

De acuerdo al criterio de conectividad, el número de conglomerados sugerido es 2; mientras que bajo el criterio de Dunn, el número sugerido es 8.

De los gráficos podemos observar que la mayoría de los criterios coinciden en que el número adecuado de conglomerados es 2, entonces se tiene la siguiente conformación:

Listing 10: Gráficos.

```
res=pam(scale(departamentos),2)
plotcluster(departamentos,res$clustering)

clusplot(departamentos,res$clustering, color = TRUE,
         shade = TRUE, labels =2,lines=0,
         main ="Grafico de Conglomerados")
```



Con respecto al agrupamiento con k-means, en el método PAM la única diferencia está en que el departamento de Lambayeque fue asignado al otro clúster.

b) Considerando una técnica de clúster jerárquico aglomerativa (AGNES) determine el enlace más apropiado y el número de conglomerados para agrupar a los departamentos. Utilice al menos cinco criterios de comparación que sustenten su respuesta. (**1.5 puntos**)

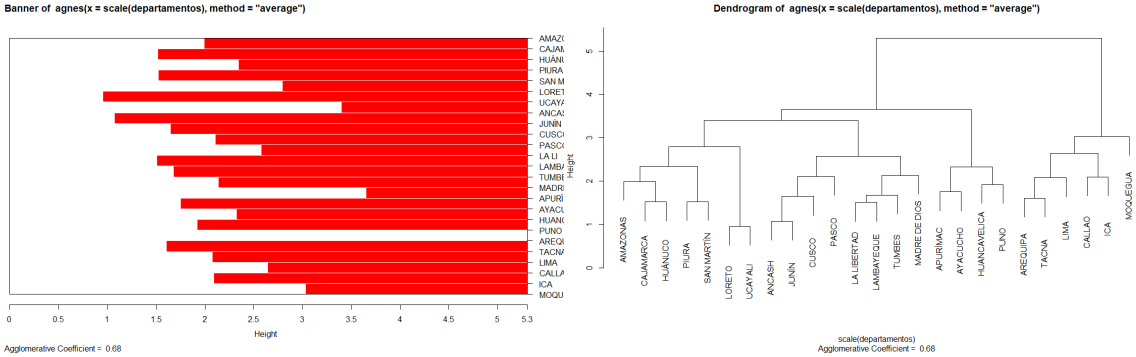
Nota: Código en **ANEXO 2**.

Para determinar el enlace más apropiado vamos a comparar los coeficientes de aglomeración (AC) obtenidos con cada uno de los enlaces. Mientras mayor sea el valor de AC (máximo es 1) se tendrá una mejor conformación de conglomerados (elementos más similares al interior y conglomerados más diferentes unos de otros). A continuación se presentan las salidas por tipo de enlace, enseguida, determinaremos el número de conglomerados para agrupar a los departamentos, en cada cuadro de los enlaces se remarca en amarillo los respectivos AC.

Listing 11: Enlace Average.

```
res=agnes(scale(departamentos),method="average")
res
```

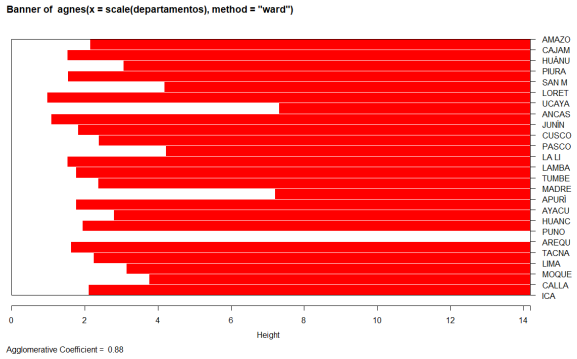
```
Call:     agnes(x = scale(departamentos), method = "average")
Agglomerative coefficient:  0.6791217
Order of objects:
 [1] AMAZONAS      CAJAMARCA     HUÁNUCO       PIURA         SAN MARTÍN    LORETO
 [7] UCAYALI       ANCASH        JUNÍN         CUSCO         PASCO         LA LIBERTAD
[13] LAMBAYEQUE    TUMBES        MADRE DE DIOS APURÍMAC      AYACUCHO      HUANCAVELICA
[19] PUNO          AREQUIPA      TACNA         LIMA          CALLAO        ICA
[25] MOQUEGUA
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9546  1.6352  2.0840  2.2347  2.5907  5.3039

Available components:
[1] "order"    "height"   "ac"       "merge"    "diss"     "call"     "method"
[8] "order.lab" "data"
```

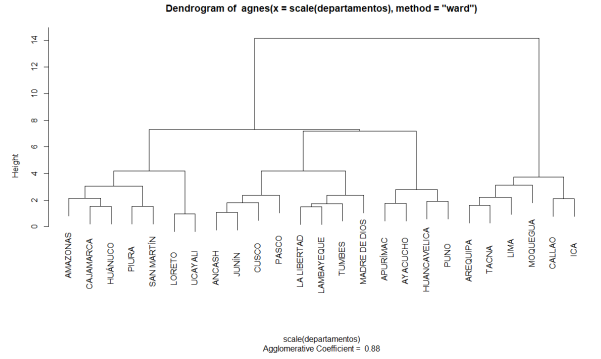


Listing 12: Enlace Ward.

```
res=agnes(scale(departamentos),method="ward")
res
```

```
Call:     agnes(x = scale(departamentos), method = "ward")
Agglomerative coefficient:  0.8769061
Order of objects:
 [1] AMAZONAS      CAJAMARCA     HUÁNUCO       PIURA         SAN MARTÍN    LORETO
 [7] UCAYALI       ANCASH        JUNÍN         CUSCO         PASCO         LA LIBERTAD
[13] LAMBAYEQUE    TUMBES        MADRE DE DIOS APURÍMAC      AYACUCHO      HUANCAVELICA
[19] PUNO          AREQUIPA      TACNA         LIMA          MOQUEGUA      CALLAO
[25] ICA
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9546  1.7073  2.1823  3.1814  3.2891 14.1874

Available components:
[1] "order"    "height"   "ac"       "merge"    "diss"     "call"     "method"
[8] "order.lab" "data"
```

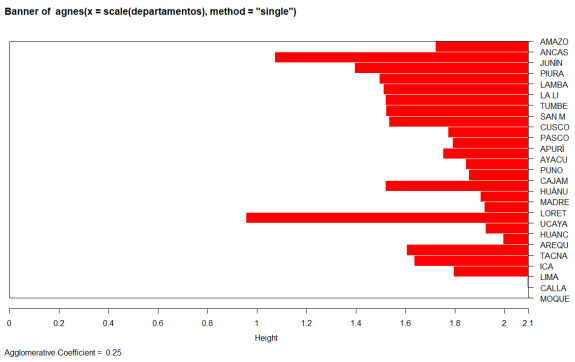


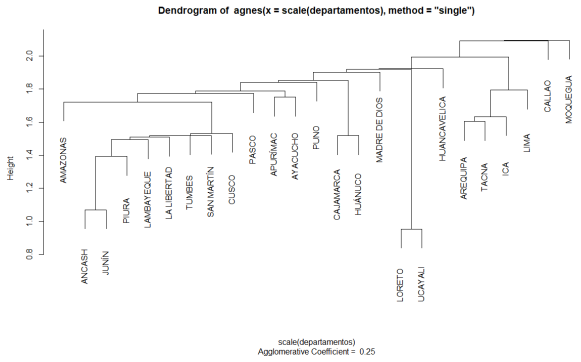


Listing 13: Enlace Simple (tiene un bajo coeficiente aglomerativo).

```
res=agnes(scale(departamentos),method="single")
res
```

```
Call:     agnes(x = scale(departamentos), method = "single")
Agglomerative coefficient:  0.2453802
Order of objects:
 [1] AMAZONAS       ANCASH         JUNÍN          PIURA          LAMBAYEQUE     LA LIBERTAD
 [7] TUMBES         SAN MARTÍN     CUSCO          PASCO          APURÍMAC       AYACUCHO
[13] PUNO           CAJAMARCA      HUÁNUCO        MADRE DE DIOS  LORETO         UCAYALI
[19] HUANCAVELICA   AREQUIPA       TACNA          ICA            LIMA           CALLAO
[25] MOQUEGUA
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9546  1.5184  1.7360  1.6748  1.8662  2.0957

Available components:
[1] "order"    "height"   "ac"       "merge"    "diss"     "call"     "method"
[8] "order.lab" "data"
```

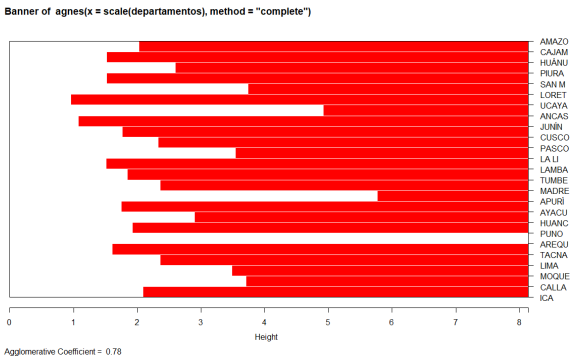


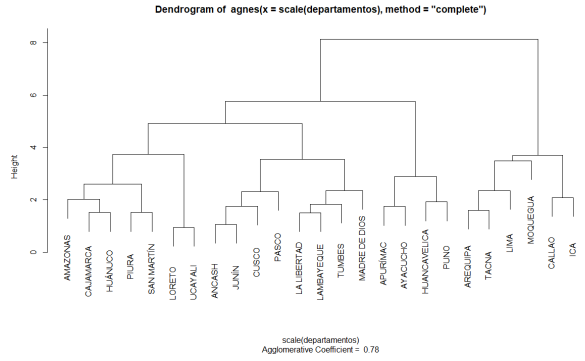


Listing 14: Enlace Completo.

```
res=agnes(scale(departamentos),method="complete")
res
```

```
Call:     agnes(x = scale(departamentos), method = "complete")
Agglomerative coefficient:  0.7835257
Order of objects:
 [1] AMAZONAS      CAJAMARCA     HUÁNUCO       PIURA         SAN MARTÍN    LORETO
 [7] UCAYALI       ANCASH        JUNÍN         CUSCO         PASCO         LA LIBERTAD
[13] LAMBAYEQUE    TUMBES        MADRE DE DIOS APURÍMAC      AYACUCHO      HUANCAVELICA
[19] PUNO          AREQUIPA      TACNA         LIMA          MOQUEGUA      CALLAO
[25] ICA
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9546  1.7143  2.2080  2.7247  3.5001  8.1362

Available components:
[1] "order"    "height"   "ac"       "merge"    "diss"     "call"     "method"
[8] "order.lab" "data"
```

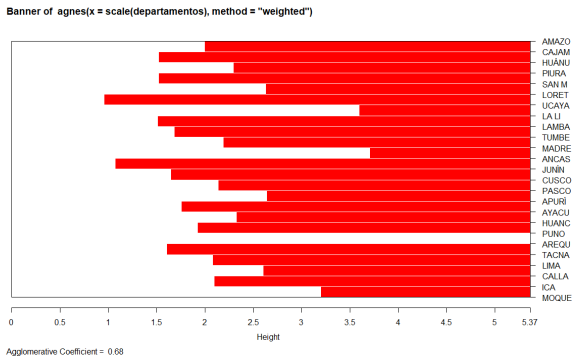


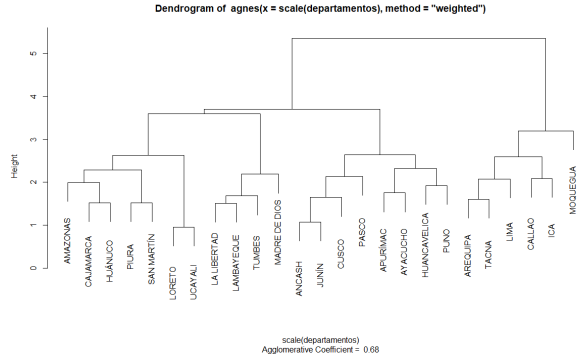


Listing 15: Enlace Ponderado.

```
res=agnes(scale(departamentos),method="weighted")
res
```

```
Call:     agnes(x = scale(departamentos), method = "weighted")
Agglomerative coefficient:  0.6809014
Order of objects:
 [1] AMAZONAS       CAJAMARCA      HUÁNUCO        PIURA          SAN MARTÍN     LORETO
 [7] UCAYALI        LA LIBERTAD    LAMBAYEQUE     TUMBES         MADRE DE DIOS  ANCASH
[13] JUNÍN          CUSCO          PASCO          APURÍMAC       AYACUCHO       HUANCAVELICA
[19] PUNO           AREQUIPA       TACNA          LIMA           CALLAO         ICA
[25] MOQUEGUA
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9546  1.6352  2.0840  2.2501  2.6061  5.3651

Available components:
[1] "order"     "height"    "ac"        "merge"     "diss"      "call"      "method"
[8] "order.lab" "data"
```

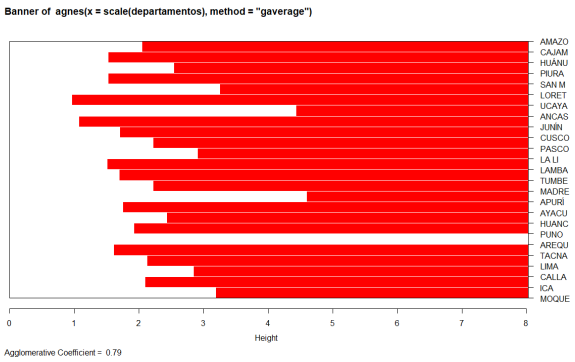


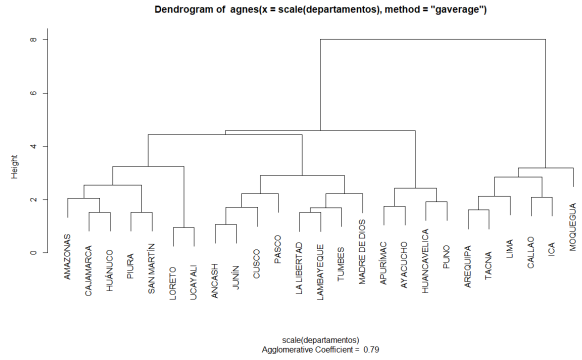


Listing 16: Enlace Ponderado Generalizado.

```
res=agnes(scale(departamentos),method="gaverage")
res
```

```
Call:     agnes(x = scale(departamentos), method = "gaverage")
Agglomerative coefficient:  0.7854132
Order of objects:
 [1] AMAZONAS       CAJAMARCA      HUÁNUCO        PIURA          SAN MARTÍN     LORETO
 [7] UCAYALI        ANCASH         JUNÍN          CUSCO          PASCO          LA LIBERTAD
[13] LAMBAYEQUE     TUMBES         MADRE DE DIOS  APURÍMAC       AYACUCHO       HUANCAVELICA
[19] PUNO           AREQUIPA       TACNA          LIMA           CALLAO         ICA
[25] MOQUEGUA
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9546  1.6729  2.1076  2.5061  2.8595  8.0279

Available components:
[1] "order"     "height"    "ac"        "merge"     "diss"      "call"      "method"
[8] "order.lab" "data"
```



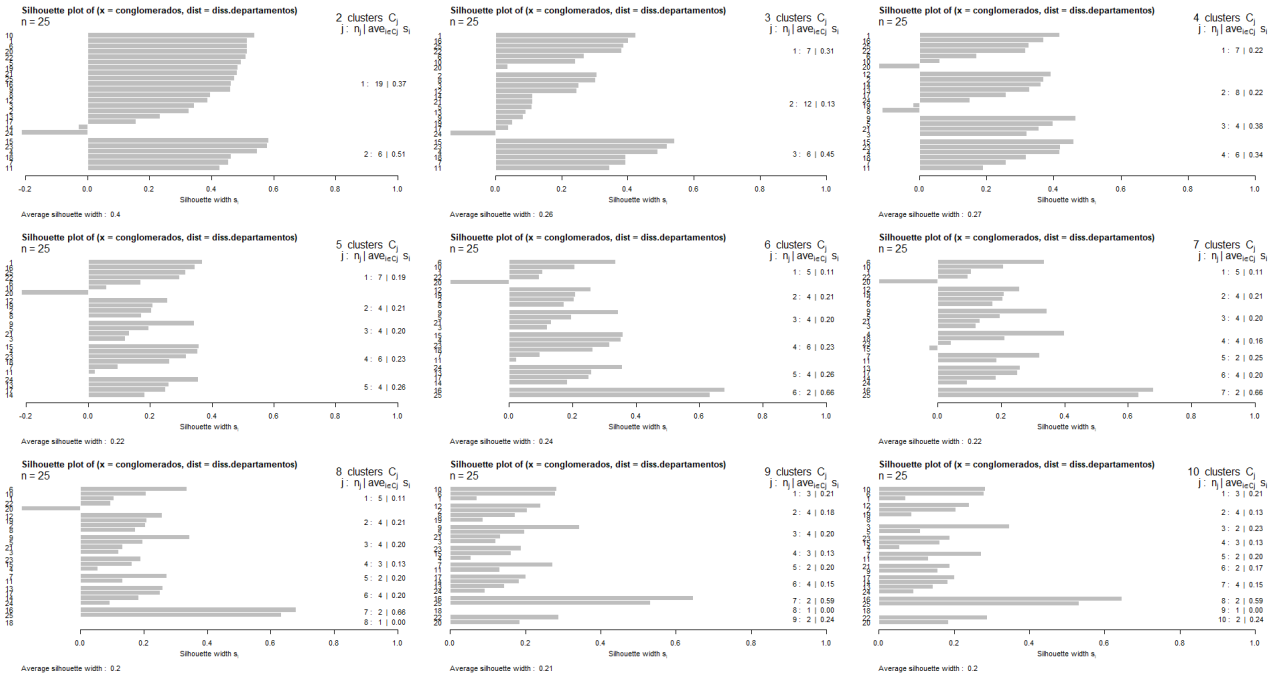


El enlace más apropiado entre los presentados en las salidas: “average”, “Ward”, “single”, “Completo”, “Ponderado”, “Ponderado Generalizado”, sería el enlace de Ward, el valor de su coeficiente de aglomeración es el mayor con 87.6 %.

Aplicamos el criterio de silueta, en la gráfica siguiente podemos observar el agrupamiento de 2 a 10 conglomerados, el promedio del índice de silueta para la primera solución con 2 conglomerados es de 0.4, siendo el mayor índice de todos.

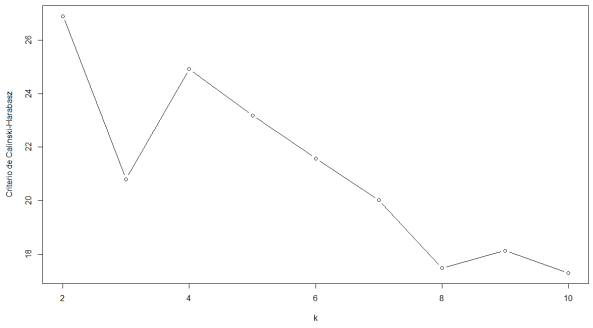
Listing 17: **Criterio 1**. Silueta.

```
diss.departamentos=daisy(scale(departamentos))
res=agnes(scale(departamentos),method="ward")

par(mfrow=c(3,3))
for(h in 2:10){
        conglomerados=cutree(res,h)
        plot(silhouette(conglomerados,diss.departamentos))
}
```



Listing 18: **Criterio 2**. Calinski-Harabasz.

```
ch<-numeric()
for(h in 2:10){
        conglomerados=cutree(res,h)
        ch<-c(ch,calinhara(diss.
            departamentos,conglomerados))
}
plot(2:10,ch,type="b",xlab="k",
ylab="Criterio de Calinski-Harabasz")
ch
```



26.90285 20.80285 24.92639 23.18923 21.57650 20.01574 17.47895 18.13029 17.28729

Listing 19: **Criterio 3 y 4**. Medidas de Validación Interna.

```
optimalScores(intern)
```

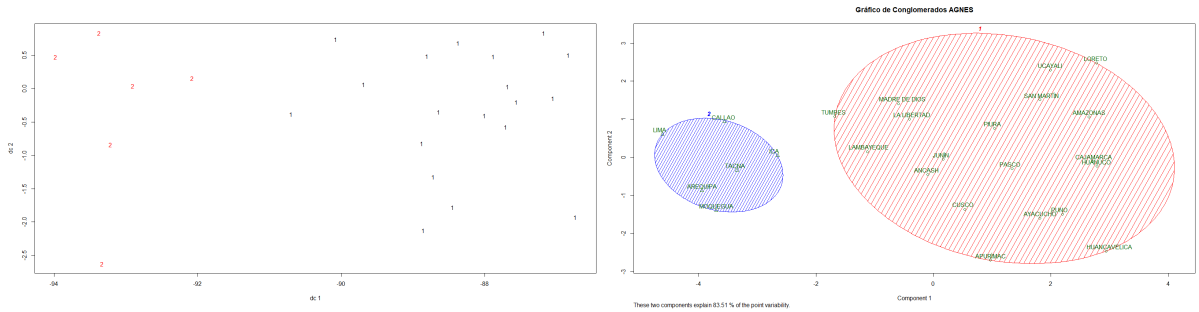
```
Optimal Scores:

             Score  Method Clusters
Connectivity 3.1667 agnes  2
Dunn         0.5905 agnes  10
Silhouette   0.4044 agnes  2
```

De acuerdo al criterio de conectividad, el número de conglomerados sugerido es 2; mientras que bajo el criterio de Dunn, el número sugerido es 10.

De los gráficos podemos observar que la mayoría de los criterios coinciden en que el número adecuado de conglomerados es 2, entonces se tiene la siguiente conformación:

Listing 20: Gráficos.

```
res_ag<-agnes(scale(departamentos),method="ward")
conglomerados_ag<-cutree(res_ag,2)
plotcluster(departamentos,conglomerados_ag)

clusplot(departamentos,conglomerados_ag, color = TRUE,  shade = TRUE, labels =2,
    lines=0,
main ="Grafico de Conglomerados AGNES")
```



Con respecto al agrupamiento con k-means: los departamentos de Lambayeque y Tumbes, fueron asignados al otro clúster. Con respecto al agrupamiento con PAM, el departamento Lambayeque fue asignado a otro clúster En el agrupamiento con AGNES el departamento de Tumbes, fue asignado al conglomerado 1 (cambió respecto al agrupamiento de PAM).

c) Considerando una técnica de cluster jerárquico divisiva (DIANA) determine el enlace más apropiado y el número de conglomerados para agrupar a los departamentos. Utilice al menos cinco criterios de comparación que sustenten su respuesta.

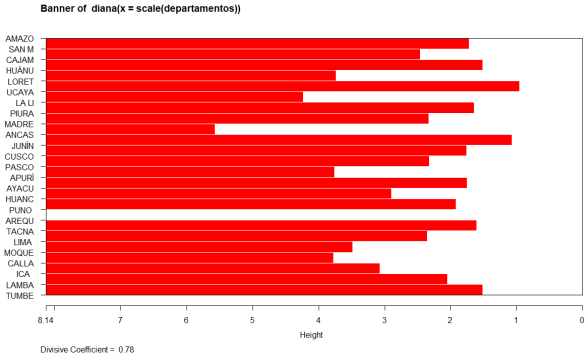
Nota: Código en **ANEXO 3**.

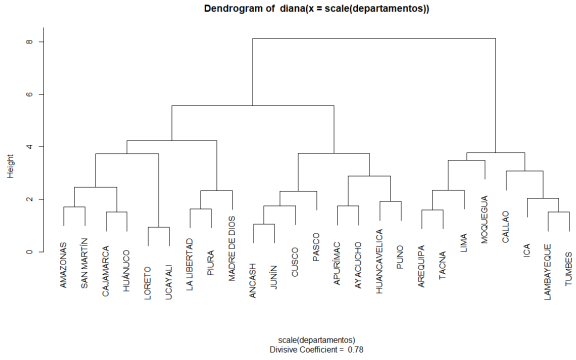
En el clúster jerárquico tenemos clústers anidados, es decir unos dentro de otros, en esta técnica jerarquica se considera que todas las observaciones están en un solo conglomerado y se van a ir dividiendo de acuerdo a su disimilaridad.

Listing 21: **Criterio 1**. Dendograma.

```
1 res=diana(scale(departamentos))
```

```
Merge:
      [,1] [,2]
 [1,]  -16  -25
 [2,]   -2  -12
 [3,]   -6  -10
 [4,]  -14  -24
 [5,]   -4  -23
 [6,]  -13  -20
 [7,]   -1  -22
 [8,]   -3   -5
 [9,]    2   -8
[10,]   -9  -21
[11,]  -11    4
[12,]    9  -19
[13,]    6  -17
[14,]    5  -15
[15,]    7    3
[16,]    8   10
[17,]   -7   11
[18,]   14  -18
[19,]   15    1
[20,]   12   16
[21,]   18   17
[22,]   19   13
[23,]   22   20
[24,]   23   21
Order of objects:
 [1] AMAZONAS      SAN MARTÍN    CAJAMARCA     HUÁNUCO       LORETO        UCAYALI
 [7] LA LIBERTAD   PIURA         MADRE DE DIOS ANCASH        JUNÍN         CUSCO
[13] PASCO         APURÍMAC      AYACUCHO      HUANCAVELICA  PUNO          AREQUIPA
[19] TACNA         LIMA          MOQUEGUA      CALLAO        ICA           LAMBAYEQUE
[25] TUMBES
Height:
 [1] 1.7208656 2.4600939 1.5181083 3.7351491 0.9546384 4.2372803 1.6475980 2.3321544
 [9] 5.5697217 1.0703545 1.7598888 2.3242103 3.7581494 1.7512116 2.8978048 1.9220979
[17] 8.1362248 1.6037108 2.3586338 3.4864522 3.7748217 3.0768577 2.0470766 1.5184712
Divisive coefficient:
[1] 0.7797534

Available components:
[1] "order"    "height"   "dc"       "merge"    "diss"     "call"     "order.lab"
[8] "data"
```



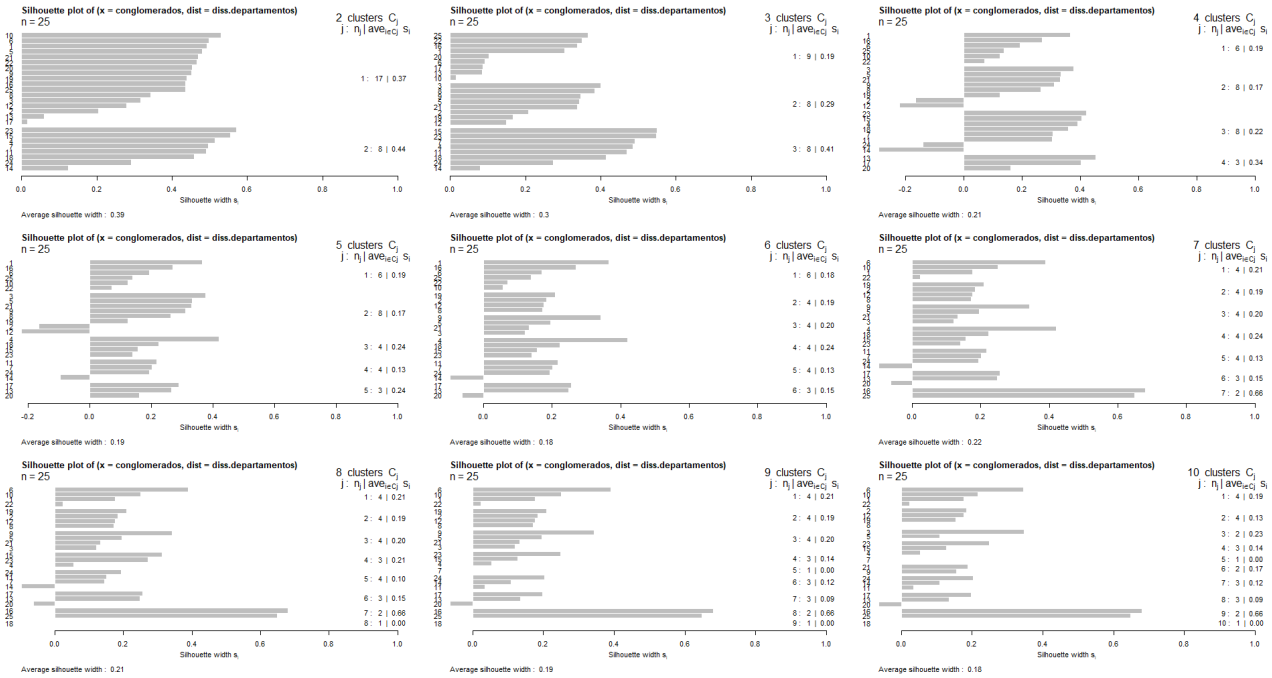


Del resultado anterior, nos podemos quedar con 2 conglomerados. A partir del dendograma tenemos las distintas conformaciones de los departamentos, de acuerdo al gráfico podríamos sugerir usar 2 conglomerados. Los departamentos que conforman cada conglomerado se puede ver en el mismo gráfico (opción cutree en R).

Aplicamos el criterio de silueta, buscando el mayor entre los valores resultantes desde 2 a 10 clusters. En la gráfica se observa que el mayor valor es 0.39 obteniendo 2 clusters.

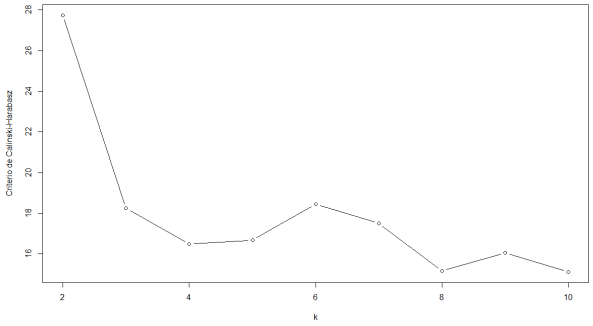
Listing 22: **Criterio 2**. Silueta.

```
diss.departamentos=daisy(scale(departamentos))
res=diana(scale(departamentos))par(mfrow=c(3,3))
for(h in 2:10){
        conglomerados=cutree(res,h)
        plot(silhouette(conglomerados,diss.departamentos))
}
```



Listing 23: **Criterio 3**. Calinski-Harabasz.

```
res=diana(scale(departamentos))
ch<-numeric()
for(h in 2:10){
        conglomerados=cutree(res,h)
        ch<-c(ch,calinhara(diss.
            departamentos,conglomerados))
}
plot(2:10,ch,type="b",xlab="k",
ylab="Criterio de Calinski-Harabasz")
ch
```



27.74361 18.25401 16.48967 16.68075 18.43856 17.50202 15.17308 16.05252 15.1060

De la salida observamos que índice de Calinski-Harabasz al parecer obtiene el mayor valor con 2 conglomerados.

Listing 24: **Criterio 4**. Medidas de Validación Interna.

```
optimalScores(intern)
```

```
Optimal Scores:

             Score  Method Clusters
Connectivity 5.2333 diana  2
Dunn         0.5661 diana  10
Silhouette   0.3949 diana  2
```

De acuerdo al criterio de conectividad, el número de conglomerados sugerido es 2; mientras que bajo el criterio de Dunn, el número sugerido es 10.

d) Compare los resultados obtenidos en las preguntas anteriores y determine cuál sería el número de conglomerados que finalmente deberían de considerarse para agrupar a los departamentos del Perú en relación a las variables descritas. Describa los grupos encontrados. (**1.0 punto**)

Nota: Código en **ANEXO 4**.

Para comparar los 4 resultados anteriores (kmeans, PAM, Agnes y Diana) podríamos usar medidas internas (conectividad, Dunn y/o silueta).

Listing 25: Medida de Validación Interna.

```
clmethods <- c("kmeans","pam","agnes","diana")
intern <- clValid(scale(departamentos), nClust = 2:10,
                  clMethods = clmethods, validation = "internal")
summary(intern)
```

```
Clustering Methods:
 kmeans pam agnes diana

Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                            2       3       4       5       6       7       8       9      10

kmeans Connectivity   9.0425 20.0369 23.0595 28.3798 31.5492 34.8710 38.0190 41.9492 46.4242
       Dunn           0.2682  0.3415  0.4072  0.4072  0.4815  0.4807  0.5156  0.4806  0.5905
       Silhouette     0.3949  0.2768  0.2962  0.2738  0.2852  0.2515  0.2239  0.2081  0.1994
pam    Connectivity   7.0730 15.5004 21.5845 28.4627 32.1083 35.3929 38.0690 39.5690 41.8524
       Dunn           0.2726  0.2638  0.4103  0.4031  0.4031  0.4050  0.4954  0.4954  0.4954
       Silhouette     0.4038  0.2690  0.2896  0.2275  0.2420  0.2110  0.1836  0.1866  0.1774
agnes  Connectivity   6.3619 14.5167 22.6437 25.9282 29.8476 33.6167 38.8063 41.9492 45.1409
       Dunn           0.3457  0.3636  0.3729  0.3729  0.3775  0.3933  0.4806  0.4806  0.5905
       Silhouette     0.4044  0.2542  0.2741  0.2267  0.2315  0.2115  0.2037  0.2081  0.2017
diana  Connectivity   9.0425 19.4754 23.6706 29.9361 34.9107 38.0802 40.6635 42.8635 46.0552
       Dunn           0.2682  0.3287  0.3690  0.3706  0.3729  0.3995  0.4527  0.4806  0.5661
       Silhouette     0.3949  0.2953  0.2115  0.1877  0.1803  0.2243  0.2072  0.1901  0.1783

Optimal Scores:

             Score  Method Clusters
Connectivity 6.3619 agnes  2
Dunn         0.5905 kmeans 10
Silhouette   0.4044 agnes  2
```

De acuerdo a la validación interna el que obtiene mejores resultados es el método AGNES con 2 conglomerados. Para las medidas externas se tiene que kmeans con 2 conglomerados es mejor.

Listing 26: Medida de estabilidad.

```
stab <- clValid(scale(departamentos), nClust = 2:10, clMethods = clmethods,
                validation = "stability")
summary(stab)
```

```
Optimal Scores:

    Score  Method Clusters
APN 0.0324 kmeans 2
AD  1.3146 kmeans 10
ADM 0.1622 kmeans 2
FOM 0.5708 diana  10
```

Por tanto el número de conglomerados necesarios para agrupar los departamentos según sus variables es 2.

En el siguiente cuadro se observa la descripción de los conglomerados encontrados, los diferentes métodos dan resultados muy parecidos.

Listing 27: Cluster.

```
head(departamentos.new)
```

	clusterkm	clusterpam	clusterdiana	clusteragnes
AMAZONAS	1	1	1	1
ANCASH	1	1	1	1
APURÍMAC	1	1	1	1
AREQUIPA	2	2	2	2
AYACUCHO	1	1	1	1
CAJAMARCA	1	1	1	1

Descripción de los conglomerados:

Listing 28: KMEANS.

```
medkm<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$
    clusterkm),FUN = mean)
medkm
```

	Group.1	vida	alfabetismo	escolaridad	ingreso	identidad	salud	saneamiento	electrificacion	policia
1	1	71.46118	88.33246	83.50694	262.3187	95.66401	10.10757	44.56979	60.25599	0.6880359
2	2	74.91875	96.38030	88.81530	427.0072	97.79536	19.16550	73.04426	83.19896	1.3016930

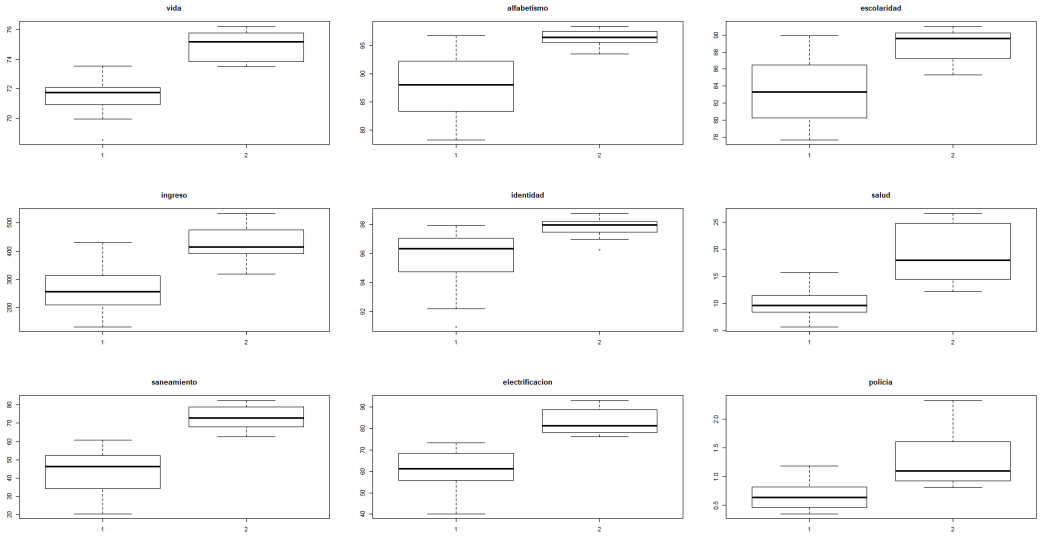


Figura 1: KMEANS

Listing 29: PAM.

```
medpam<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$
    clusterpam),FUN = mean)
medpam
```

```
  Group.1     vida alfabetismo escolaridad  ingreso identidad    salud saneamiento
1       1 71.57667    88.61826    83.60809 265.4330  95.73732 10.36774    45.78454
2       2 75.11571    96.79509    89.31353 442.5258  97.91131 19.79048    73.98841
  electrificacion   policia
1        61.13506 0.7066367
2        84.21607 1.3415276
```

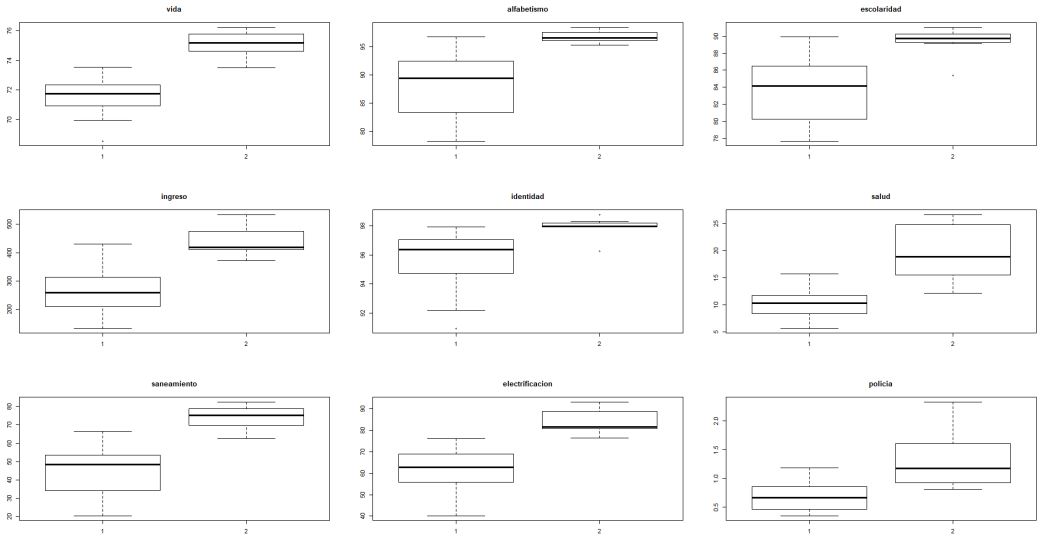


Figura 2: PAM

Listing 30: AGNES.

```
med_ag<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$
    clusteragnes),FUN = mean)
med_ag
```

```
  Group.1     vida alfabetismo escolaridad  ingreso identidad    salud saneamiento
1       1 71.76737    89.03805    83.70128 273.1905  95.76474 10.46320    46.67082
2       2 75.10167    96.82856    89.96932 447.4758  98.18682 21.05867    75.88250
  electrificacion   policia
1         62.1872 0.7203829
2         84.7311 1.4038131
```

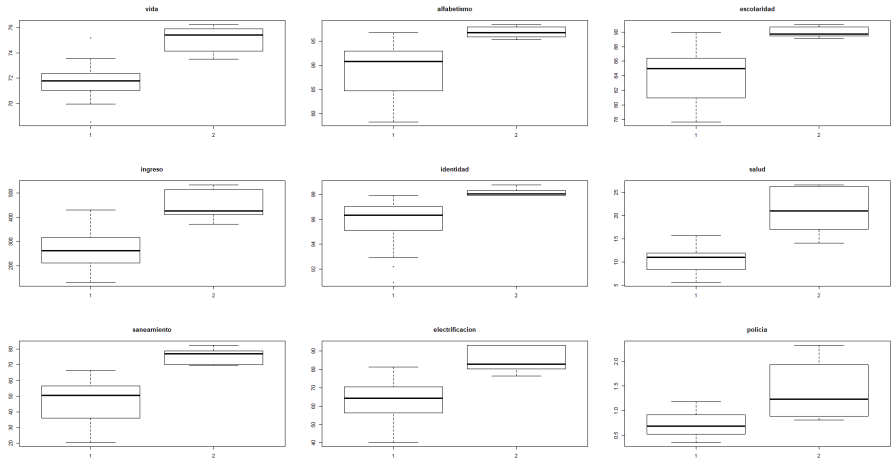


Figura 3: AGNES

Listing 31: DIANA.

```
1 med_di<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$
    clusterdiana),FUN = mean)
2 med_di
```

```
  Group.1     vida alfabetismo escolaridad  ingreso identidad    salud saneamiento
1       1 71.76737    89.03805    83.70128 273.1905  95.76474 10.46320    46.67082
2       2 75.10167    96.82856    89.96932 447.4758  98.18682 21.05867    75.88250
  electrificacion   policia
1         62.1872 0.7203829
2         84.7311 1.4038131
```

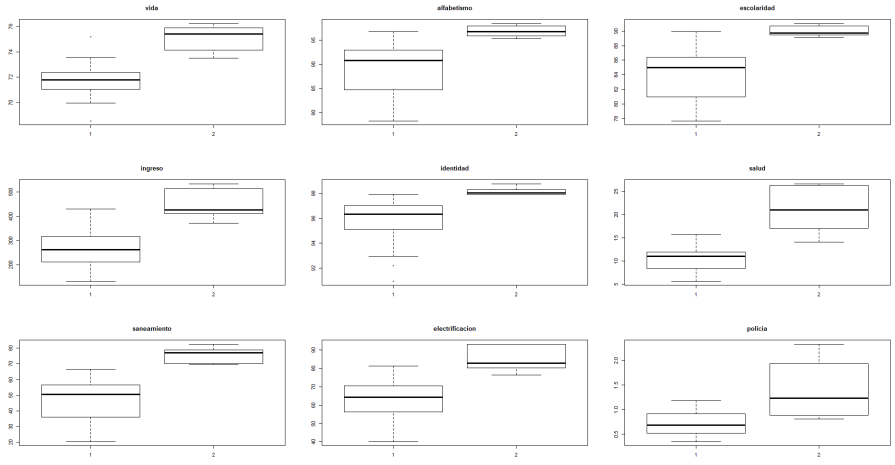


Figura 4: DIANA

Para los 4 métodos se observa lo siguiente:

El cluster 1 está conformado por departamentos con menor esperanza de vida al nacer (aprox. 72 años) en comparación con el cluster 2 (aprox. 75 años).

En el cluster 1 se tiene menor tasa de alfabetismo en adultos en promedio (aprox. 89 %) con respecto al cluster 2 (aprox. 97 %).

Los departamentos que conforman el cluster 1 tienen un ingreso familiar mensual en promedio mucho menor con respecto al cluster 2.

En el cluster 1 se tiene una menor tasa de escolaridad en promedio (aprox. 85 %) con respecto al cluster 2 (aprox. 90 %).

Los departamentos que están en el clúster 1 tienen una menor población con acta de nacimiento (aprox. 96 %) con respecto al clúster 2 (aprox. 98 %) En el cluster 1 el promedio de médicos es menor (11 médicos aprox. por cada 10,000 habitantes) con respecto al cluster 2 (20 médicos aprox. por cada 10,000 habitantes).

Los departamentos en el cluster 1 tienen en promedio menos viviendas con agua y desagüe (aprox. 50 %) con respecto al cluster 2 (aprox. 73 %). En el cluster 1 se encuentran los departamentos donde hay en promedio menos viviendas con electrificación (aprox. 64En el cluster 1 están los departamentos donde hay menos policías en promedio (0.7 policías aprox. por cada mil habitantes) con respecto al cluster 2 (1.3 policías aprox. por cada mil habitantes).

Según la descripción de variables, el primer conglomerado lo conforman los departamentos con los indicadores más bajos en promedio (pobres): Amazonas, Ancash, Apurímac, Ayacucho, Cajamarca, Cusco, Huancavelica, Huánuco, Junín, La Libertad, Lambayeque (según el método agnes), Loreto, Madre de Dios, Pasco, Piura, Puno, San Martin, Ucayali, Tumbes (según el método agnes). Mientras que el segundo conglomerado está conformado por Arequipa, Callao, Ica, Lima, Moquegua, Tacna.

PREGUNTA 2 (7 puntos)

NOTA: Código en **ANEXO 5**

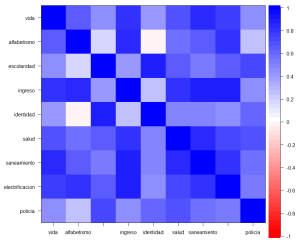
La técnica de análisis factorial requiere la verificación de supuestos para que su uso sea el apropiado; en primer lugar, verificaremos la correlación entre par de variables en la base de datos y luego analizaremos la correlación a nivel multivariado mediante la prueba de Kaiser-Meyer-Olkin (KMO). Además, comprobaremos que, si el modelo factorial es adecuado, entonces la mayoría de los elementos no diagonales de la matriz Anti-imagen deberían ser pequeños.

En las siguientes imágenes se verifica que casi todas las variables (dos a dos) están correlacionadas, encontrándose solo débil asociación entre las variables identidad vs alfabetismo, escolaridad vs alfabetismo y policía vs alfabetismo. Esto marcaría un primer indicio de que es conveniente usar la técnica de análisis factorial.

Listing 32: Matriz y gráfico de correlaciones

```
R = round(cor(departamentos),3)
R
```

	vida	alfabetismo	escolaridad	ingreso	identidad	salud	saneamiento	electrificacion	policia
vida	1.000	0.634	0.415	0.790	0.376	0.664	0.833	0.727	0.449
alfabetismo	0.634	1.000	0.141	0.830	-0.057	0.545	0.638	0.790	0.250
escolaridad	0.415	0.141	1.000	0.411	0.858	0.627	0.518	0.609	0.725
ingreso	0.790	0.830	0.411	1.000	0.246	0.769	0.876	0.868	0.442
identidad	0.376	-0.057	0.858	0.246	1.000	0.488	0.457	0.426	0.582
salud	0.664	0.545	0.627	0.769	0.488	1.000	0.826	0.699	0.649
saneamiento	0.833	0.638	0.518	0.876	0.457	0.826	1.000	0.779	0.548
electrificacion	0.727	0.790	0.609	0.868	0.426	0.699	0.779	1.000	0.518
policia	0.449	0.250	0.725	0.442	0.582	0.649	0.548	0.518	1.000



Antes de aplicar la prueba de KMO, realizaremos la prueba de Esfericidad de Barlett, planteando la hipótesis de que la matriz de correlaciones de la muestra es igual a una matriz identidad; obligando que, para hacer esta prueba de Esfericidad, se debe verificar previamente normalidad multivariada en la matriz de datos original. De este modo, asumiendo normalidad en la base de datos "departamentos", el test de Barlett nos da un P-valor inferior al 5 % por lo que se rechaza la hipótesis planteada y evidenciando que las variables de la matriz de datos están correlacionadas.
Ahora mediante el test de Kaiser-Meyer-Olkin, obtenemos un valor de 0.836, lo que nos indica una fuerte asociación entre las variables, por lo que sería apropiado usar la técnica de análisis factorial. Así mismo, revisando la matriz de correlación Anti-imagen, se observa en la parte superior de la diagonal que la mayoría de sus valores son pequeños (< 0.5), indicando de esta manera que es adecuado usar el análisis factorial. Por otro lado, si calculamos la Medida de Adecuación Muestral (MSA) en todas las variables, éstas tendrían valores cercanos a la unidad (> 0.7), por lo que se sugiere utilizar a todas las variables, ya que aportan buena información al momento de usar la técnica factorial.

Listing 33: Matriz de correlación Anti-imagen.

```
round(descri$Anti.Image.Cor,3)
```

```
             vida alfabetismo escolaridad ingreso identidad  salud saneamiento electrificacion policia
vida        1.000      -0.098       0.126  -0.108    -0.187  0.102      -0.376          -0.042  -0.066
alfabetismo -0.098       1.000       0.297  -0.233     0.263 -0.174       0.135          -0.628  -0.088
escolaridad  0.126       0.297       1.000  -0.038    -0.623 -0.305       0.242          -0.491  -0.412
ingreso     -0.108      -0.233      -0.038   1.000     0.242 -0.236      -0.487          -0.343   0.116
identidad   -0.187       0.263      -0.623   0.242     1.000  0.038      -0.308          -0.110   0.125
salud        0.102      -0.174      -0.305  -0.236     0.038  1.000      -0.359           0.276  -0.196
saneamiento -0.376       0.135       0.242  -0.487    -0.308 -0.359       1.000          -0.028  -0.117
electrificacion -0.042   -0.628      -0.491  -0.343    -0.110  0.276      -0.028           1.000   0.087
policia     -0.066      -0.088      -0.412   0.116     0.125 -0.196      -0.117           0.087   1.000
```

Listing 34: Medidas individuales de Adecuación Muestral.

```
t(round(descri$MSA,3))
```

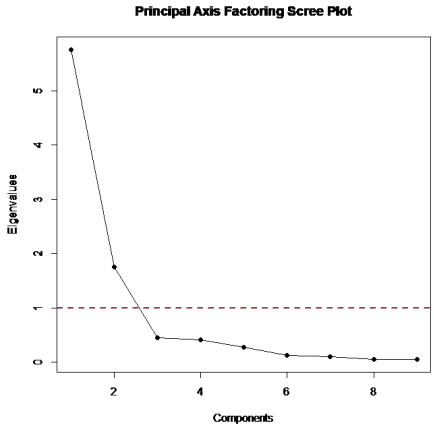
```
     vida alfabetismo escolaridad ingreso identidad salud saneamiento electrificacion policia
MSA 0.933       0.788       0.715   0.875     0.739 0.891        0.85           0.818   0.896
```

Selección del número de factores:

El gráfico de Sedimentación nos ayudará a elegir el número de factores que retengan la mayor información, en este caso, a partir del tercer factor la pendiente formada por los autovalores se estabiliza, por lo que se decide hacer el análisis factorial solo con dos factores, de otra manera, si tomamos en cuenta la regla de Kaiser que nos sugiere tomar tantos factores como autovalores mayores a la unidad, optaríamos también con dos factores. Otra técnica que se puede usar para la elección de factores es mediante el análisis paralelo de la libraría en R "pysch" con el método de máxima verosimilitud, lo que nos recomienda utilizar también solo dos factores.

Listing 35: Selección del número de Factores

```
scree(departamentos)
fa.parallel(departamentos, fm="ml",
            fa="fa")
```




```
Parallel analysis suggests that the number of factors =  2
and the number of components =  NA
```

Comunalidades y Cargas factoriales:

La solución con dos factores, implica que el modelo factorial explique el 79.3 % de la varianza total y para saber si las variables están bien representadas por los factores, se deben calcular las comunalidades y éstas deben tener valores superiores a 0.4 (sugeridas por los autores). En esta aplicación el valor mínimo de la comunalidad la obtiene la variable "policía" (0.55986) lo cual implica que todas las variables están bien representadas por los dos factores.

Listing 36: Comunalidades.

```
factanal.none = factanal(departamentos, factors=2, rotation="none")
comunal = 1 - factanal.none$uniquenesses
comunal
```

```
          vida     alfabetismo     escolaridad
       0.67271         0.79711         0.90310
       ingreso       identidad          salud
       0.96553         0.84441         0.72957
saneamiento electrificacion         policia
       0.82937         0.83664         0.55986
```

Después de conocer que las variables están debidamente representadas por los factores, se debe saber con cuál de los dos factores se relacionan mejor las variables, para ello analizaremos las cargas factoriales. Las cargas indican el grado de correspondencia entre la variable y el Factor, es decir, que cargas altas indican que dicha variable es representativa para dicho factor.

Listing 37: Cargas Factoriales.

```
factanal.vari = factanal(departamentos, factors=2, rotatio="varimax")
factanal.none$loadings
factanal.vari$loadings
```

```
Loadings:
               Factor1 Factor2
vida            0.818
alfabetismo     0.765  -0.460
escolaridad     0.605   0.733
ingreso         0.956  -0.228
identidad       0.450   0.801
salud           0.838   0.163
saneamiento     0.911
electrificacion 0.914
policia         0.584   0.468

               Factor1 Factor2
SS loadings      5.447   1.691
Proportion Var   0.605   0.188
Cumulative Var   0.605   0.793
```

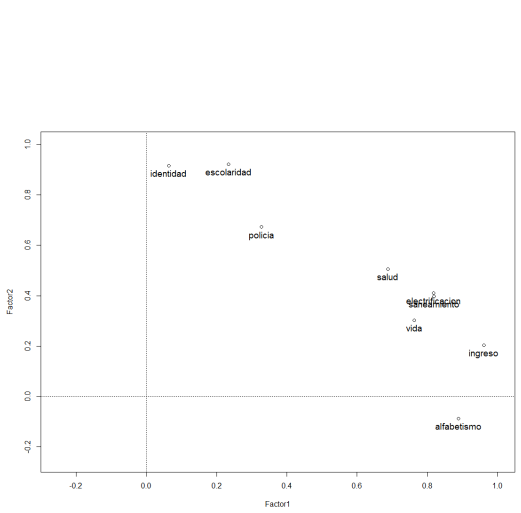
Figura 5: Sin rotación

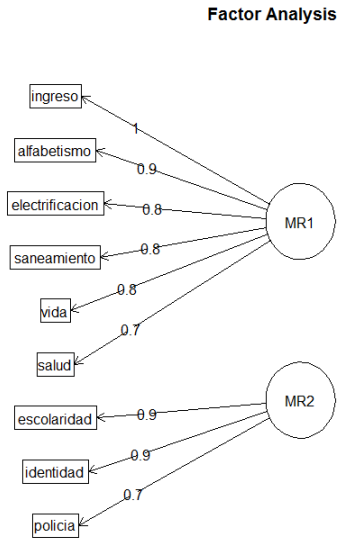
```
Loadings:
               Factor1 Factor2
vida            0.762   0.302
alfabetismo     0.888
escolaridad     0.234   0.921
ingreso         0.962   0.202
identidad               0.917
salud           0.688   0.506
saneamiento     0.819   0.398
electrificacion 0.818   0.410
policia         0.328   0.673

               Factor1 Factor2
SS loadings      4.274   2.864
Proportion Var   0.475   0.318
Cumulative Var   0.475   0.793
```

Figura 6: Con rotación, *Varimax*

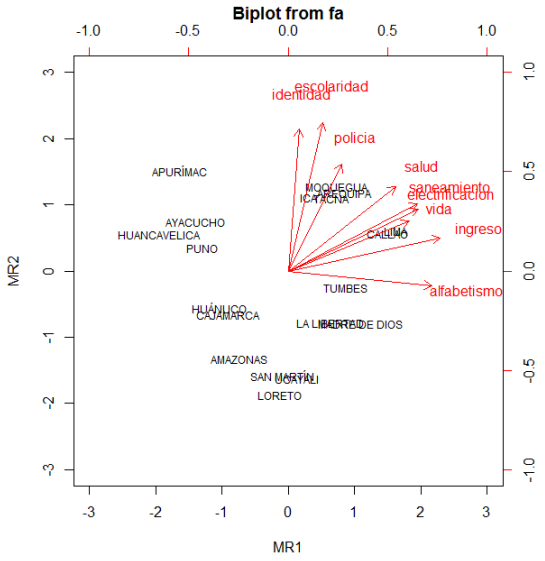
El propósito de la rotación es conseguir que las cargas factoriales sea lo más próximo a la unidad con solo uno de los factores. Así, por ejemplo, para la variable "escolaridad" se tiene una carga factorial de 0.605 y 0.733 con el factor 1 y factor 2 respectivamente, lo cual no nos ayudaría a representar mejor la variable en cada factor ya que disponen de cargas factoriales similares. Por ello, se realiza una rotación denominada "Varimax" para que la variable tenga mejor representación en solo uno de los factores; de esta manera, la variable escolaridad esta mejor representada con el factor 2, pues su carga factorial luego de la rotación es de 0.921.

Finalmente, las variables vida, alfabetismo, ingreso, salud, saneamiento y electrificación serian mejor explicadas en el factor 1, mientras que las variables escolaridad, identidad y policía se explicarían en el factor 2. De este modo, hemos reducido las 9 variables originales a 2 factores para el estudio de algunas características de la población por departamento.

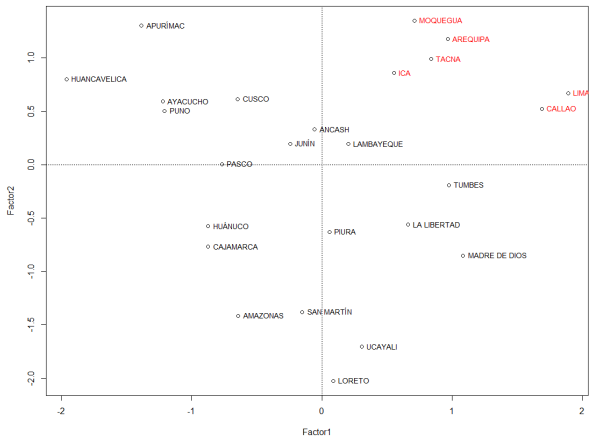




En el gráfico "biplot" se aprecia 3 bloques de atributos relacionados, siendo el primer bloque conformado por los atributos de identidad, escolaridad y policía, el segundo bloque lo conformarían los atributos de salud, saneamiento, electrificación, vida e ingreso; y el tercer bloque conformado solo por el atributo de alfabetismo. El primer bloque coincide con las variables del segundo factor, mientras que el segundo y tercer bloque lo tiene el primer factor. En el gráfico también se observa que los departamentos de Moquegua, Arequipa, Tacna, Ica, Lima y la provincia del Callao disponen de una mayor cantidad de policías por cada habitante y una alta tasa de escolaridad (5 a 18 años) en relación al resto de departamentos. Por otra parte, Lima y Callao son los que mejores condiciones de vida poseen como el de salud, saneamiento, viviendas con electricidad, ingreso familiar y esperanza de vida. Los departamentos más notorios que no gozarían de buenas condiciones de vida son Cajamarca, Huánuco, Amazonas y San Martin.



En el siguiente gráfico según la técnica de análisis factorial, se aprecia que en el primer cuadrante se ubican los departamentos en mejores condiciones como son el de salud, ingreso familiar per cápita y esperanza de vida, mientras que el resto de departamentos no gozarían de buenas condiciones. Este resultado es similar a los grupos formados en el análisis de conglomerados desarrollado en la anterior pregunta, puesto que se forman dos grupos de departamentos las cuales uno de ellos se ubica en el primer cuadrante (Moquegua, Arequipa, Tacna, Ica, Lima y Callao).



PREGUNTA 3 (6 puntos)

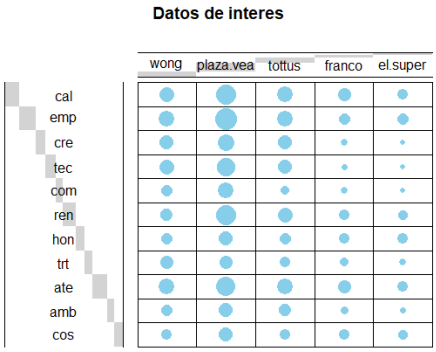
Para el siguiente conjunto de datos es de interés evaluar la percepciónn de los residentes de una ciudad sobre los distintos supermercados que operan en ella. Para ello se tomó una muestra aleatoria simple de 401 residentes de esta ciudad entre los 18 y 70 años edad de todos los niveles socioeconómicos. Se realizará un análisis de correspondencia.

NOTA: Código en **ANEXO 6**

Listing 38: Tabla de Contingencia.

```
wong=c(113,126,99,119,64,79,67,87,121,64,55)
plaza.vea=c(207,244,137,167,124,205,95,87,190,97,99)
tottus=c(122,126,97,106,41,114,67,56,122,72,54)
franco=c(87,66,27,18,28,61,56,39,85,33,59)
el.super=c(53,62,16,15,16,51,53,24,55,21,51)

M=as.table(cbind(wong,plaza.vea,tottus,franco,el.super))
rownames(M)=c("cal","emp","cre","tec","com","ren","hon","trt",
"ate","amb","cos")
M
```

```
    wong plaza.vea tottus franco el.super
cal  113       207    122     87       53
emp  126       244    126     66       62
cre   99       137     97     27       16
tec  119       167    106     18       15
com   64       124     41     28       16
ren   79       205    114     61       51
hon   67        95     67     56       53
trt   87        87     56     39       24
ate  121       190    122     85       55
amb   64        97     72     33       21
cos   55        99     54     59       51
```



- Las filas están determinadas como: **cal:** Ofrece productos de calidad, **emp:** Genera empleo, **cre:** Creativa e innovadora, **tec:** Tiene tecnología moderna, **com:** Ayuda a la comunidad, **ren:** Es rentable, **hon:** Es honesta/paga impuestos, **trt:** Paga bien/trata bien a empleados, **ate:** Brinda buena atención a clientes, **amb:** Protege el medio ambiente, **cos:** Respeta costumbres de pobladores.
- Como logra observarse, las mayores frecuencias se encuentran en el supermercado PLAZA VEA, sin embargo el análisis de correspondencia analizará que dato (o datos) de interés exclusivo lo caracteriza, y que datos de interés sobresalen más en otros.

Listing 39: Prueba χ^2 de Pearson.

```
chisq.test(M)
```

```
        Pearson's Chi-squared test

data:  M
X-squared = 196.26, df = 40, p-value < 2.2e-16
```

- La prueba establece la hipótesis nula de independencia entre las filas (datos de interés) y las columnas (supermercados). Dado que obtenemos un $p - valor < 0.05$ rechazamos tal hipótesis nula; y por ende, establecemos que existe dependencia entre los datos de interés y los supermercados. Por lo que tendría sentido el análisis de correspondencia.

Se realizarán tablas de probabilidades condicionales,

Listing 40: Probabilidad condicional bajo un dato de interés fijo.

```
prop.table(M, 1)
```

	wong	plaza.vea	tottus	franco	el.super
cal	0.19415808	0.35567010	0.20962199	0.14948454	0.09106529
emp	0.20192308	0.39102564	0.20192308	0.10576923	0.09935897
cre	0.26329787	0.36436170	0.25797872	0.07180851	0.04255319
tec	0.28000000	0.39294118	0.24941176	0.04235294	0.03529412
com	0.23443223	0.45421245	0.15018315	0.10256410	0.05860806
ren	0.15490196	0.40196078	0.22352941	0.11960784	0.10000000
hon	0.19822485	0.28106509	0.19822485	0.16568047	0.15680473
trt	0.29692833	0.29692833	0.19112628	0.13310580	0.08191126
ate	0.21116928	0.33158813	0.21291449	0.14834206	0.09598604
amb	0.22299652	0.33797909	0.25087108	0.11498258	0.07317073
cos	0.17295597	0.31132075	0.16981132	0.18553459	0.16037736

Listing 41: Probabilidad condicional bajo un tipo de supermercado.

```
prop.table(M, 2)
```

	wong	plaza.vea	tottus	franco	el.super
cal	0.11368209	0.12530266	0.12487206	0.15563506	0.12709832
emp	0.12676056	0.14769976	0.12896622	0.11806798	0.14868106
cre	0.09959759	0.08292978	0.09928352	0.04830054	0.03836930
tec	0.11971831	0.10108959	0.10849539	0.03220036	0.03597122
com	0.06438632	0.07506053	0.04196520	0.05008945	0.03836930
ren	0.07947686	0.12409201	0.11668373	0.10912343	0.12230216
hon	0.06740443	0.05750605	0.06857728	0.10017889	0.12709832
trt	0.08752515	0.05266344	0.05731832	0.06976744	0.05755396
ate	0.12173038	0.11501211	0.12487206	0.15205725	0.13189448
amb	0.06438632	0.05871671	0.07369498	0.05903399	0.05035971
cos	0.05533199	0.05992736	0.05527124	0.10554562	0.12230216

- Con respecto a la primera tabla, observamos que las mayores probabilidades condicionales se encuentran bajo el supermercado Plaza Vea, esto quiere decir que si partimos de un dato de interés fijo, es más probable de que la opinión establecida indique que Plaza Vea sea la mejor. Es claro que solo para el dato de interés en el que se paga o trata bien a los empleados, indiquen a mejores supermercados a Wong y Plaza Vea.

- Con respecto a la segunda tabla, el análisis es totalmente distinto. Por ejemplo, Si partimos del supermercado Franco, entonces puede indicarse que un dato de interés en el que sobrevale dicho supermercado es en el de atención a clientes, pues como logra observase tiene una mayor probabilidad condicional de tal dato de interés a diferencia de los otros supermercados.

Estos tipos de análisis son los que se desean estudiar para establecer correspondencias adecuadas.

A continuación, se generan datos de un análisis de correspondecias resumido.

Listing 42: Análisis de correspondencias (1)

```
fit <- ca(M)
print(fit)
summary(fit)
```

```
Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.030752  72.1  72.1  ******************
 2      0.006905  16.2  88.2  ****
 3      0.003383   7.9  96.2  **
 4      0.001635   3.8 100.0  *
        -------- -----
 Total: 0.042675 100.0


Rows:
     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1  |  cal |  127  557   26 |  -68 532  19 |   15  25   4 |
2  |  emp |  136  664   23 |    6   4   0 |   69 660  93 |
3  |  cre |   82  962  126 |  247 923 162 |  -51  39  30 |
4  |  tec |   92  981  247 |  333 975 334 |  -25   6   9 |
5  |  com |   59  453   82 |  125 265  30 |  106 188  96 |
6  |  ren |  111  878   62 |  -41  71   6 |  139 807 311 |
7  |  hon |   73  941  144 | -269 863 173 |  -81  78  70 |
8  |  trt |   64  845   67 |   26  15   1 | -193 829 345 |
9  |  ate |  125  785   25 |  -72 612  21 |  -38 173  26 |
10 |  amb |   62  461   18 |   63 324   8 |  -41 137  15 |
11 |  cos |   69  986  180 | -330 985 245 |  -12   1   1 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | wong |  216  924  180 |  134 502 125 | -122 421 468 |
2 | plzv |  359  943  128 |   70 322  57 |   97 621 491 |
3 | ttts |  212  358   90 |   79 341  43 |  -18  18  10 |
4 | frnc |  122  930  282 | -301 913 357 |  -42  17  30 |
5 | elsp |   91  940  320 | -376 940 418 |    4   0   0 |
```

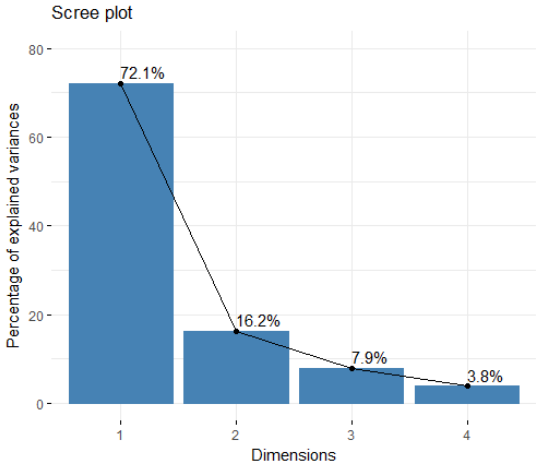
- Como se observa en los resultados, los 2 mayores eigenvalores establecen 88.2 % de información acumulada por los datos, por lo que sería suficiente tomar dos dimensiones y tratar de explicar los datos en base a ellas.
- En los resultados podemos observar **qlt:** indicando la calidad de representacion de los datos por parte de las dos primeras dimensiones establecidas. Como se logra observa, los datos de interés **com** y **amb** son quienes poseen una baja calidad de representación por parte de las dos dimensiones.

Se realizarán análisis gráficos y descriptivos, usando la librería FactoMineR.

Listing 43: Eigenvalores y Porcentaje de variabilidad

```
eig.val <- get_eigenvalue(res.ca)
eig.val
fviz_screeplot(res.ca, addlabels = TRUE, ylim = c(0, 80))
```

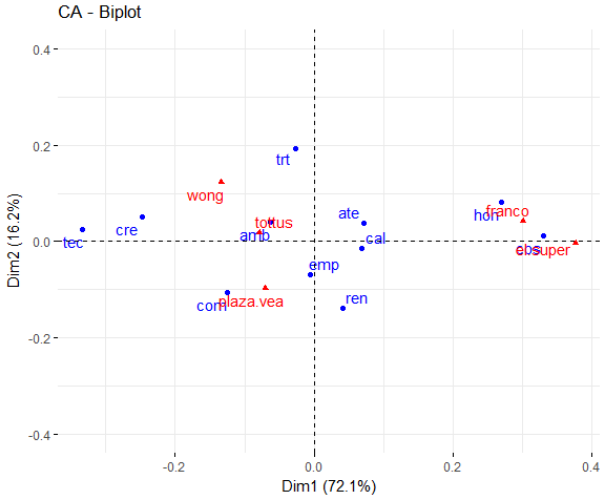
```
      eigenvalue variance.percent cumulative.variance.percent
Dim.1 0.030751788        72.060266                    72.06027
Dim.2 0.006905298        16.181094                    88.24136
Dim.3 0.003383393         7.928261                    96.16962
Dim.4 0.001634618         3.830379                   100.00000
```



- Se logra observar lo anteriormente dicho, las dos primeras dimensiones manejan el 88.2 % de la variabilidad de los datos.

Listing 44: Análisis de correspondencias (1)

```
        fit <- ca(M)
        print(fit)
        summary(fit)
```



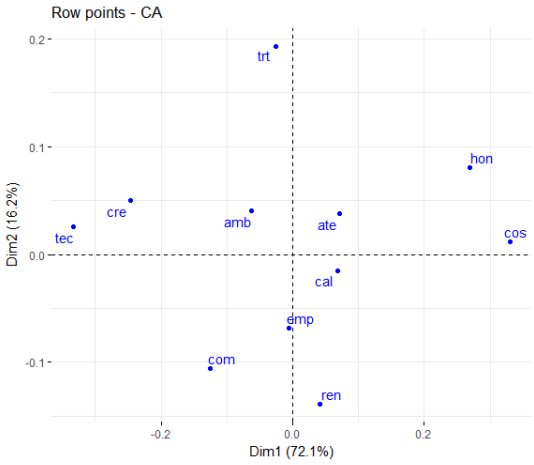
- Podemos decir según la gráfica que los datos de interés de tecnólogía y creatividad están relacionados.

- El datos de interés **amb** puede que esté más relacionado al supermercado Tottus, así como **hon** al supermercado Franco, **emp** y **com** relacionados con el supermercado **Plaza vea**.
- Es cuestión de analizar si están bien representados, lo cual se verá a continuación.

Listing 45: Coordenadas de los datos de interés

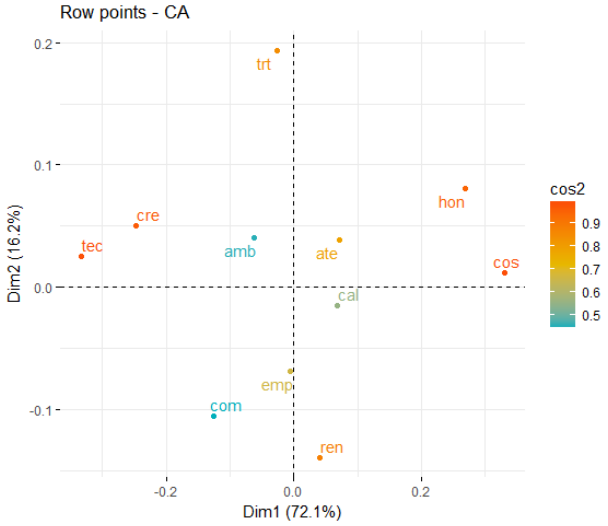
```
head(row$coord)
fviz_ca_row(res.ca, repel = TRUE)
```

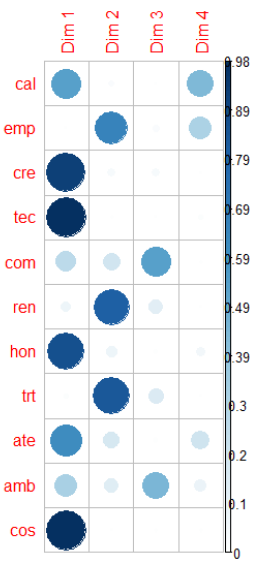
```
> head(row$coord)
           Dim 1       Dim 2       Dim 3       Dim 4
cal  0.068143865 -0.01483475  0.003580483 -0.062069960
emp -0.005572736 -0.06870755 -0.014508699  0.046811964
cre -0.246910874  0.05057372  0.049719945 -0.004946540
tec -0.333271498  0.02548304  0.015759410  0.043923831
com -0.125313380 -0.10552926 -0.179011640 -0.018384455
ren  0.041288057 -0.13915381  0.054115298 -0.002910645
```

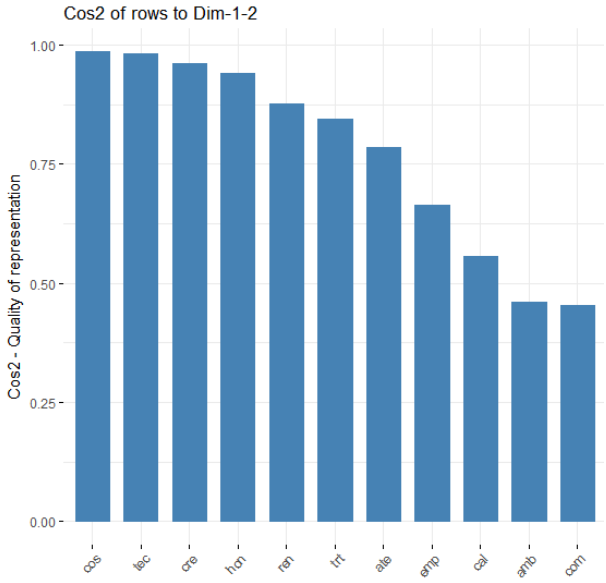


Listing 46: Asociación de filas sobre dimensiones

```
fviz_ca_row(res.ca, col.row = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
library("corrplot")
corrplot(row$cos2, is.corr=FALSE)
fviz_cos2(res.ca, choice = "row", axes = 1:2)
```



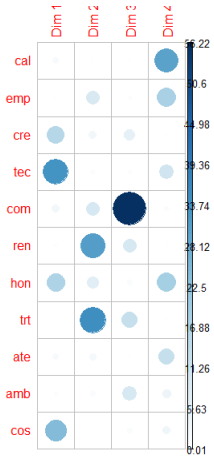


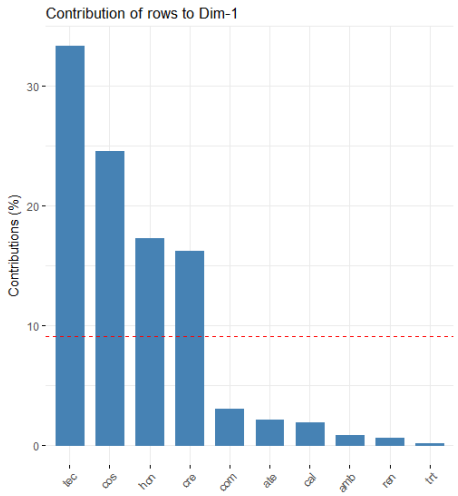


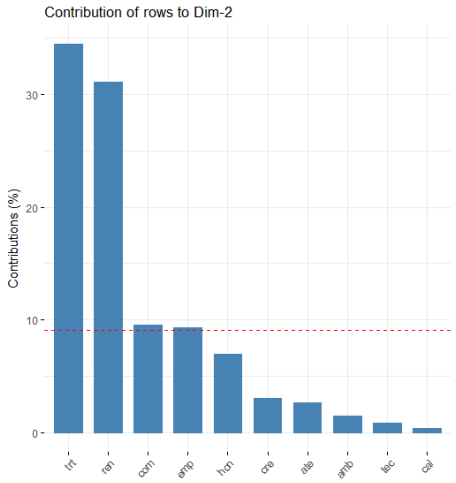
- Como se observa, los datos de interés **amb** y **com** son quienes poseen una baja representación por parte de las dimensiones tomadas, a diferencia de las demás que parecen estar relativamente bien explicadas.

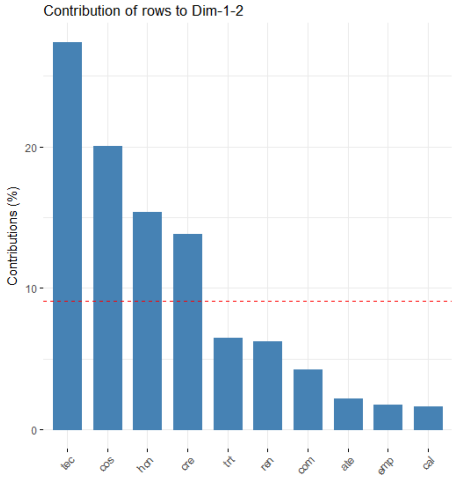
Listing 47: Contribución de los datos de interés sobre las dimensiones

```
corrplot(row$contrib, is.corr=FALSE)
# Contribuciones de filas a la dimension 1
fviz_contrib(res.ca, choice = "row", axes = 1, top = 10)
# Contribuciones de filas a la dimension 2
fviz_contrib(res.ca, choice = "row", axes = 2, top = 10)
# Contribucion total
fviz_contrib(res.ca, choice = "row", axes = 1:2, top = 10)
```







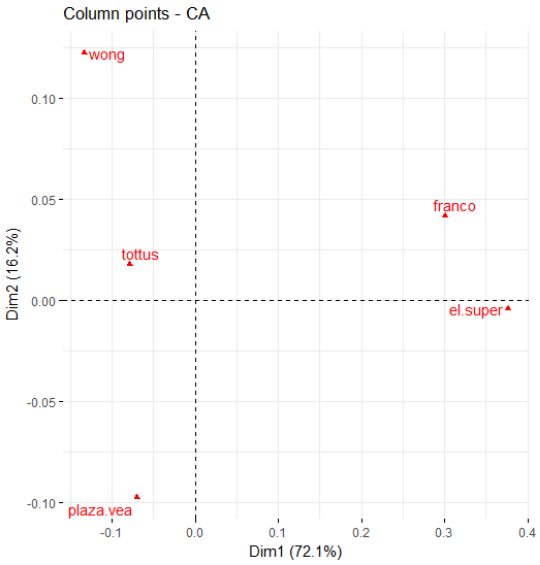


- Se observa que los datos de interés de tecnología, costumbres, honestidad y creatividad son cuales más contribuyen a la definición de la dimensión 1.
- Se observa que los datos de interés de buen trato, rentabilidad, ayuda a la comunidad y empleo son cuales más contribuyen a la definición de la dimensión 2.
- Si embargo se observa que el dato de interés de empleo queda relegada muy ampliamente, lo cual es debido a que si bien define buena parte de la dimensión 2, contribuye mucho más en la definición de una dimensión no tomada en el análisis.

Listing 48: Coordenadas de los supermercados

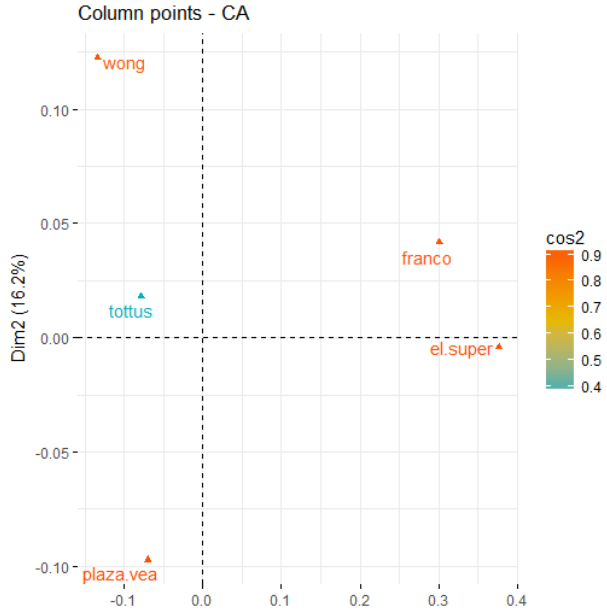
```
head(col$coord)
fviz_ca_col(res.ca, repel = TRUE)
```

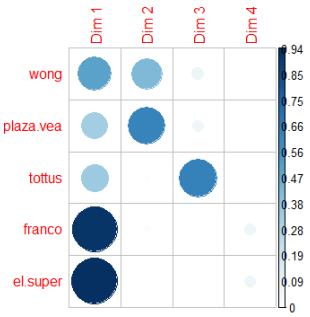
```
                Dim 1       Dim 2       Dim 3       Dim 4
wong      -0.13350752  0.122284020 -0.04936844  0.016237472
plaza.vea -0.06995789 -0.097200565 -0.02941304 -0.001167596
tottus    -0.07853993  0.017847534  0.10742521 -0.009015014
franco     0.30065312  0.041610561 -0.02449523 -0.079429994
el.super   0.37636771 -0.004010375  0.01535004  0.093520040
```

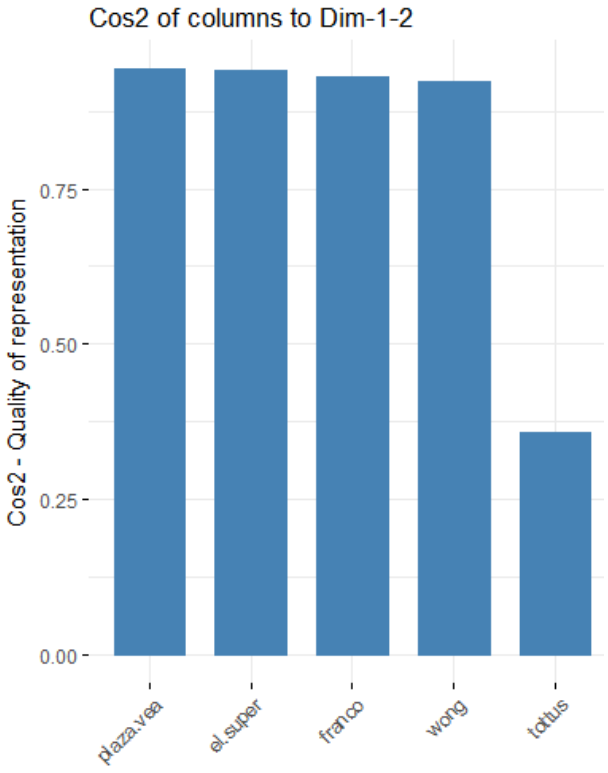


Listing 49: Asociación de columnas (Supermercados) sobre dimensiones

```
fviz_ca_col(res.ca, col.col = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
library("corrplot")
corrplot(col$cos2, is.corr=FALSE)
fviz_cos2(res.ca, choice = "col", axes = 1:2)
```



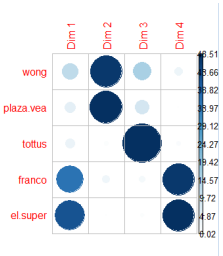


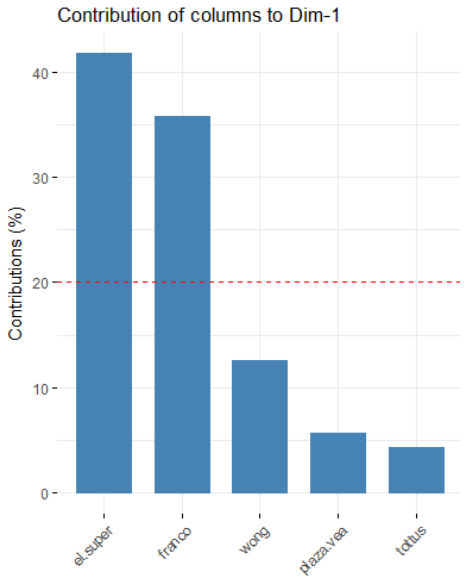


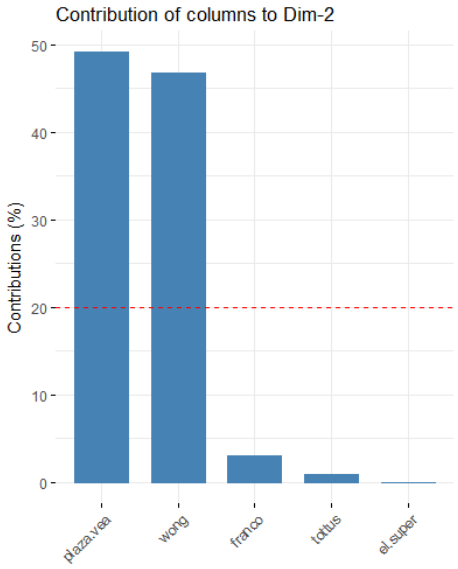
- Como se observa, el supermercado **Tottus** es quien posee una baja representación por parte de las dimensiones tomadas, a diferencia de las demás que parecen estar relativamente bien explicadas.

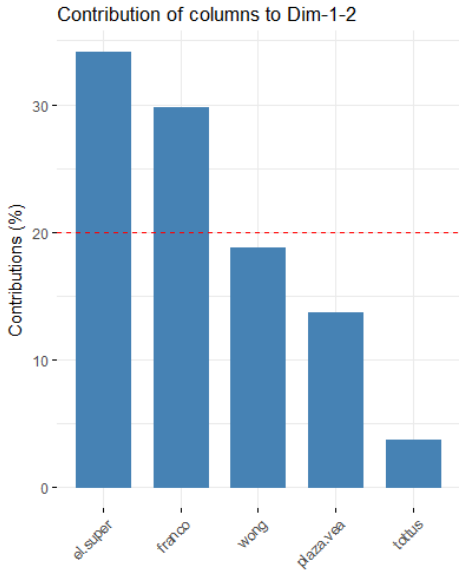
Listing 50: Contribución de los supermercados sobre las dimensiones

```
corrplot(col$contrib, is.corr=FALSE)
# Contribuciones de columnas a la dimension 1
fviz_contrib(res.ca, choice = "col", axes = 1, top = 10)
# Contribuciones de columnas a la dimension 2
fviz_contrib(res.ca, choice = "col", axes = 2, top = 10)
# Contribucion total
fviz_contrib(res.ca, choice = "col", axes = 1:2, top = 10)
fviz_ca_col(res.ca, col.col = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
```





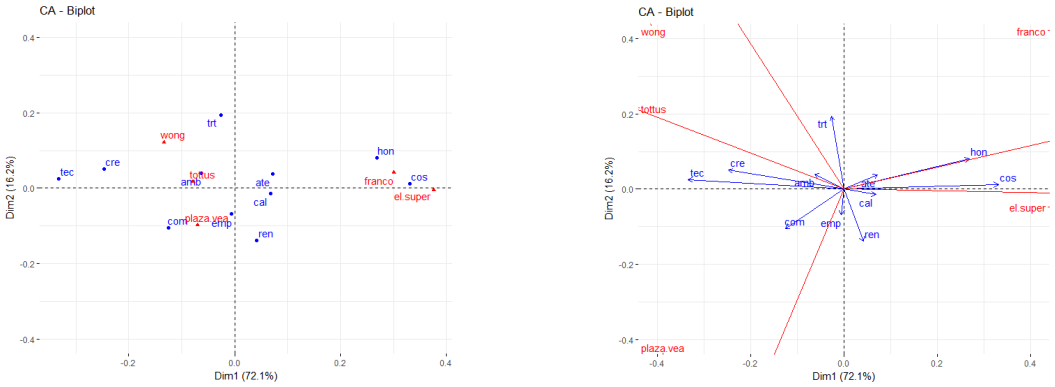




- Se observa que los supermercados Wong y Plaza Vea son cuales más contribuyen a la definición de la dimensión 1.
- Se observa que los supermercados Franco y Super son cuales más contribuyen a la definición de la dimensión 2, sin embargo contribuyen en mayor medida la definición de las dimensiones que no se están optando.
- Se observa que los supermercados Franco y Super contribuyen en mayor totalidad a la definición de las dimensiones por lo que la variabilidad del conjunto de datos viene explicada en mayor parte por dichos supermercados.

Listing 51: Biplot de asociaciones

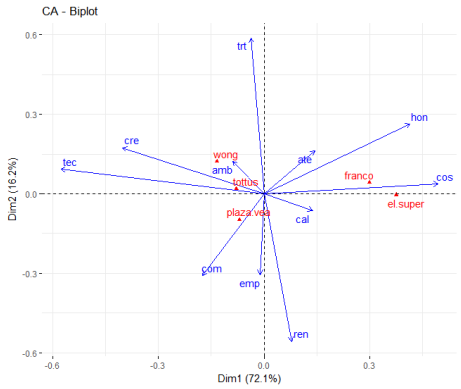
```r
# Simetrico
fviz_ca_biplot(res.ca, repel = TRUE,ylim=c(-0.4,0.4))

# Asimetrico
fviz_ca_biplot(res.ca,
map ="rowprincipal", arrow = c(TRUE, TRUE),
repel = TRUE,xlim=c(-0.4,0.4),ylim=c(-0.4,0.4))
```



- Existe una alta asociación entre el dato de interés de honestidad y paga impuestos con respecto al supermercado Franco. Así como el buen trato también posee una alta asociación con el supermercado Wong. Y por último, el respeto a las costumbres asociado a El Super.

Listing 52: Biplot de contribuciones

```r
fviz_ca_biplot(res.ca, map ="colgreen", arrow = c(TRUE, FALSE),
repel = TRUE)
```



- Existen mejores contribuciones en la definición de las dimensiones por partes de los datos de interés de tecnología, creatividad, buen trato, honestidad, respetar costumbres y rentabilidad.

Anexo 1

Listing 53: Contribución de los supermercados sobre las dimensiones

```
### Pregunta 1 ###

rm(list = ls())

## Pregunta 1a ##

library(foreign)
library(cluster)
library(fpc)

departamentos=read.spss(file.choose(),
use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)

colnames(departamentos) <- tolower(colnames(departamentos))
nombres=departamentos[,1]
departamentos=departamentos[,-1]
rownames(departamentos)=nombres
head(departamentos)


## K means

## Determinar numero de conglomerados

# Criterio 1: Suma de cuadrados dentro de clusters

# Se crea este vector wss para ir guardando la suma de cuadrados dentro del cluster

wss<-numeric()
for(h in 2:10){
b<-kmeans(scale(departamentos),h,nstart = 20)
wss[h-1]<-b$tot.withinss
}
plot(2:10,wss,type="b")


# Criterio 2: Silueta

diss.departamentos=daisy(scale(departamentos))
par(mfrow=c(3,3))
for(h in 2:10){
res=kmeans(scale(departamentos),h)
plot(silhouette(res$cluster,diss.departamentos))
}
par(mfrow=c(1,1))

kmeansruns(scale(departamentos),criterion="asw")

# Criterio 3: de Calinski-Harabasz

ch<-numeric()
for(h in 2:10){
res<-kmeans(scale(departamentos),h, nstart = 20)
ch[h-1]<-calinhara(scale(departamentos),res$cluster)
}
```

```
plot(2:10,ch,type="b",xlab="k",
ylab="Criterio de Calinski-Harabasz")

kmeansruns(scale(departamentos),criterion="ch")


# Criterio 4 y 5: Medidas de Validacion Interna (Conectividad y Dunn)

library(clValid)
clmethods <- c("kmeans")

intern <- clValid(scale(departamentos), nClust = 2:10,
clMethods = clmethods, validation = "internal", neighbSize=5)

summary(intern)
plot(intern)
optimalScores(intern)


# Grafico

reskm=kmeans(scale(departamentos),2)
plotcluster(departamentos,reskm$cluster)

clusplot(departamentos,reskm$cluster, color = TRUE,
shade = TRUE, labels =2,lines=0,
main ="Grafico de Conglomerados")


## Metodo PAM


# Criterio 1: Suma de cuadrados dentro de cluster

asw<-numeric()
for(h in 2:10){
res<-pam(scale(departamentos),h)
asw[h-1]<-res$silinfo$avg.width
}
plot(2:10,asw,type="b",xlab="k",ylab="ASW")


# Criterio 2: Silueta
par(mfrow=c(3,3))
for(h in 2:10){
res=pam(scale(departamentos),h)
plot(res,which.plots=2)
}

pamk(scale(departamentos),criterion="asw")


# Criterio 3: de Calinski-Harabasz

par(mfrow=c(1,1))
ch<-numeric()
for(h in 2:10){
res<-pam(scale(departamentos),h)
ch[h-1]<-calinhara(scale(departamentos),res$clustering)
```

```
}
plot(2:10,ch,type="b",xlab="k",
ylab="Criterio de Calinski-Harabasz")

pamk(scale(departamentos),criterion="ch")


# Criterio 4: Medidas de Validacion Interna (Conectividad y Dunn)

library(clValid)
clmethods <- c("pam")

intern <- clValid(scale(departamentos), nClust = 2:10,
clMethods = clmethods, validation = "internal", neighbSize=5)

summary(intern)
plot(intern)
optimalScores(intern)


# Grafico

respam=pam(scale(departamentos),2)
plotcluster(departamentos,respam$clustering)
clusplot(departamentos,respam$clustering, color = TRUE,
shade = TRUE, labels =2,lines=0,
main ="Grafico de Conglomerados")
```

Anexo 2

Listing 54: Contribución de los supermercados sobre las dimensiones

```
## Pregunta 1b ##

#Clustering jerarquico aglomerativo usando Agnes

library(cluster)

# Usando Enlace promedio:

res=agnes(scale(departamentos),method="average")
res
plot(res)

# Usando Enlace de Ward:

res=agnes(scale(departamentos),method="ward")
res
plot(res)

# Usando Enlace Simple:

res=agnes(scale(departamentos),method="single")
res
plot(res)

# Usando Enlace Completo:

res=agnes(scale(departamentos),method="complete")
res
plot(res)

# Usando Enlace Ponderado:

res=agnes(scale(departamentos),method="weighted")
res
plot(res)

# Usando Enlace promedio generalizado:

res=agnes(scale(departamentos),method="gaverage")
res
plot(res)


# Obtener numero de conglomerados

# Criterio 1: Silueta:

diss.departamentos=daisy(scale(departamentos))
res=agnes(scale(departamentos),method="ward")

par(mfrow=c(3,3))
for(h in 2:10){
conglomerados=cutree(res,h)
plot(silhouette(conglomerados,diss.departamentos))
}
```

```
56
57 # Criterio 2: Calinski-Harabasz
58
59 diss.departamentos=daisy(scale(departamentos))
60 res=agnes(scale(departamentos),method="ward")
61
62 ch<-numeric()
63 for(h in 2:10){
64 conglomerados=cutree(res,h)
65 ch<-c(ch,calinhara(diss.departamentos,conglomerados))
66 }
67 plot(2:10,ch,type="b",xlab="k",
68 ylab="Criterio de Calinski-Harabasz")
69
70
71 # Criterio 3: Medidas de Validacion Interna
72
73 library(clValid)
74 clmethods <- c("agnes")
75
76 intern <- clValid(scale(departamentos), nClust = 2:10,
77 clMethods = clmethods, validation = "internal", neighbSize=5)
78
79 summary(intern)
80 plot(intern)
81 optimalScores(intern)
82
83 # Criterio 4: Medidas de estabilidad
84
85 stab <- clValid(scale(departamentos), nClust = 2:10, clMethods = clmethods,
86 validation = "stability")
87 summary(stab)
88
89 # Grafico
90 res_ag<-agnes(scale(departamentos),method="ward")
91 conglomerados_ag<-cutree(res_ag,2)
92 plotcluster(departamentos,conglomerados_ag)
93 clusplot(departamentos,conglomerados_ag, color = TRUE,  shade = TRUE, labels =2,lines
    =0,
94 main ="Grafico de Conglomerados AGNES")
```

Anexo 3

Listing 55: Contribución de los supermercados sobre las dimensiones

```
## Pregunta 1c ##

res=diana(scale(departamentos))
res
plot(res)

# Criterio 2: Silueta

diss.departamentos=daisy(scale(departamentos))
res=diana(scale(departamentos))par(mfrow=c(3,3))
for(h in 2:10){
conglomerados=cutree(res,h)
plot(silhouette(conglomerados,diss.departamentos))
}

# Criterio 3: Calinski-Harabasz

diss.departamentos=daisy(scale(departamentos))
res=diana(scale(departamentos))

ch<-numeric()
for(h in 2:10){
conglomerados=cutree(res,h)
ch<-c(ch,calinhara(diss.departamentos,conglomerados))
}
plot(2:10,ch,type="b",xlab="k",
ylab="Criterio de Calinski-Harabasz")


# Criterio 4 y 5: Medidas de Validacion Interna

library(clValid)
clmethods <- c("diana")

intern <- clValid(scale(departamentos), nClust = 2:10,
clMethods = clmethods, validation = "internal", neighbSize=5)

summary(intern)
plot(intern)
optimalScores(intern)

# Grafico

res_di<-diana(scale(departamentos))
conglomerados_di<-cutree(res_di,2)
plotcluster(departamentos,conglomerados_di)
clusplot(departamentos,conglomerados_di, color = TRUE,  shade = TRUE, labels =2,lines
    =0,
main ="Grafico de Conglomerados DIANA")
```

Anexo 4

Listing 56: Contribución de los supermercados sobre las dimensiones

```
## Pregunta 1d ##

library(clValid)
clmethods <- c("kmeans","pam","agnes","diana")

# Medidas de validacion interna
intern <- clValid(scale(departamentos), nClust = 2:10,
clMethods = clmethods, validation = "internal")
summary(intern)
plot(intern)

# Medidas de estabilidad (q tan estable es cuando se saca 1 columna cada vez

stab <- clValid(scale(departamentos), nClust = 2:10, clMethods = clmethods,
validation = "stability")
summary(stab)


#  Perfilado y caracterizacion de clusters

# Adicionar los cluster a la base de datos
departamentos.new<-cbind(departamentos,reskm$cluster)
colnames(departamentos.new)<-c(colnames(departamentos.new[,-length(departamentos.new)
    ]), "clusterkm")
head(departamentos.new)

departamentos.new<-cbind(departamentos.new,respam$cluster)
colnames(departamentos.new)<-c(colnames(departamentos.new[,-length(departamentos.new)
    ]), "clusterpam")
head(departamentos.new)

departamentos.new<-cbind(departamentos.new,conglomerados_ag)
colnames(departamentos.new)<-c(colnames(departamentos.new[,-length(departamentos.new)
    ]), "clusteragnes")
head(departamentos.new)

departamentos.new<-cbind(departamentos.new,conglomerados_di)
colnames(departamentos.new)<-c(colnames(departamentos.new[,-length(departamentos.new)
    ]), "clusterdiana")
head(departamentos.new)

# Tabla de medias
medkm<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$clusterkm),
    FUN = mean)
medkm
medpam<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$clusterpam),
    FUN = mean)
medpam
med_ag<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$clusteragnes
    ),FUN = mean)
med_ag
med_di<-aggregate(x = departamentos.new[,1:9],by = list(departamentos.new$clusterdiana
    ),FUN = mean)
med_di

```

```
48
49 # Describir variables
50
51 par(mfrow=c(3,3))
52 for (i in 1:length(departamentos.new[,1:9])) {
53 boxplot(departamentos.new[,i]~departamentos.new$clusterkm, main=names(departamentos.
       new[i]), type="l")
54 }
55
56 par(mfrow=c(1,1))
57 par(mfrow=c(3,3))
58 for (i in 1:length(departamentos.new[,1:9])) {
59 boxplot(departamentos.new[,i]~departamentos.new$clusterpam, main=names(departamentos.
       new[i]), type="l")
60 }
61
62 par(mfrow=c(1,1))
63 par(mfrow=c(3,3))
64 for (i in 1:length(departamentos.new[,1:9])) {
65 boxplot(departamentos.new[,i]~departamentos.new$clusterag, main=names(departamentos.
       new[i]), type="l")
66 }
67
68
69 par(mfrow=c(1,1))
70 par(mfrow=c(3,3))
71 for (i in 1:length(departamentos.new[,1:9])) {
72 boxplot(departamentos.new[,i]~departamentos.new$clusterdiana, main=names(departamentos
       .new[i]), type="l")
73 }
```

Anexo 5

Listing 57: Contribución de los supermercados sobre las dimensiones

```
### Cargar base de datos: "DepartamentosPeru.sav"

library(foreign)
departamentos = read.spss("DepartamentosPeru.sav",
use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
colnames(departamentos) = tolower(colnames(departamentos))
nombres = departamentos[,1]
departamentos = departamentos[,-1]
rownames(departamentos) = nombres
head(departamentos)

### Matriz y grafico de correlaciones
library(psych)
cor.plot(cor(departamentos))
R = round(cor(departamentos),3)

### Prueba de Esfericidad de Barlett
describe(departamentos)
cortest.bartlett(R,nrow(departamentos))

### Prueba de KMO
library(rela)
descri = paf(as.matrix(departamentos))
descri$KMO

### Matriz de correlacion Anti-imagen
round(descri$Anti.Image.Cor,3)

### Medidas individuales de Adecuacion Muestral
t(round(descri$MSA,3))

### Seleccion del numero de Factores
scree(departamentos)
fa.parallel(departamentos, fm="ml", fa="fa")

### Comunalidades
factanal.none = factanal(departamentos, factors=2, rotation="none")
comunal = 1 - factanal.none$uniquenesses
comunal

### Cargas Factoriales con rotacion Varimax
factanal.vari = factanal(departamentos, factors=2, rotatio="varimax")
factanal.none$loadings
factanal.vari$loadings

### Grafica de Factores
load = factanal.vari$loadings[,1:2]
plot(load, ylim=c(-0.25,1), xlim=c(-0.25,1))
abline(h=0, v=0, lty=3)
text(load, labels=colnames(departamentos), cex=1.2, pos=1)
X11()
fa.vari = fa(departamentos, nfactors=2, rotate="varimax")
fa.diagram(fa.vari, e.size=0.1)

### Grafico de Puntuaciones y Biplot
```

```
punt = factanal(departamentos, factors=2, rotation="varimax", scores="regression")
plot(punt$scores)
abline(h=0, v=0, lty=3)
text(punt$scores, labels=nombres, cex=0.9, pos=4, col=departamentos.new$clusteragnes)
X11()
biplot(fa.vari, labels=rownames(departamentos))
```

Anexo 6

Listing 58: Contribución de los supermercados sobre las dimensiones

```
library("gplots")
library(ca)

## DESARROLLO DE TABLA DE CONTINGENCIA

wong=c(113,126,99,119,64,79,67,87,121,64,55)
plaza.vea=c(207,244,137,167,124,205,95,87,190,97,99)
tottus=c(122,126,97,106,41,114,67,56,122,72,54)
franco=c(87,66,27,18,28,61,56,39,85,33,59)
el.super=c(53,62,16,15,16,51,53,24,55,21,51)

M=as.table(cbind(wong,plaza.vea,tottus,franco,el.super))
rownames(M)=c("cal","emp","cre","tec","com","ren","hon","trt",
"ate","amb","cos")
M

balloonplot(t(M), main ="Datos de interes", xlab ="", ylab="",
label = FALSE, show.margins = FALSE)

# Test chi-cuadrado
chisq <- chisq.test(M)
chisq

## ANALISIS DE CORRESPONDENCIA (1)

# Perfiles de Fila
prop.table(M, 1)

# Perfiles de Columna
prop.table(M, 2)

fit <- ca(M)
print(fit)
summary(fit)


# Grafico simetrico
plot(fit)
# Grafico asimetrico
plot(fit, mass = TRUE, contrib = "absolute", map =
"rowgreen", arrows = c(FALSE, TRUE)) # asymmetric map

## Usando FactoMineR
library(FactoMineR)
res.ca <- CA(M, graph = FALSE)

# autovalores/varianzas
library("factoextra")
eig.val <- get_eigenvalue(res.ca)
eig.val
fviz_screeplot(res.ca, addlabels = TRUE, ylim = c(0, 80))

# Biplot

fviz_ca_biplot(res.ca, repel = TRUE,ylim=c(-0.4,0.4))
```

```r

# Analisis de Filas
row <- get_ca_row(res.ca)

# coordenadas
head(row$coord)
fviz_ca_row(res.ca, repel = TRUE)
# Filas con perfiles similares son agrupadas



# Cos2: Asociacion de filas con dimensiones
head(row$cos2)
fviz_ca_row(res.ca, col.row = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
library("corrplot")
corrplot(row$cos2, is.corr=FALSE)
fviz_cos2(res.ca, choice = "row", axes = 1:2)


# Contribuciones a las dimensiones
head(row$contrib)

# para explicar la variabilidad del conjunto de datos
library("corrplot")
corrplot(row$contrib, is.corr=FALSE)
# Contribuciones de filas a la dimension 1
fviz_contrib(res.ca, choice = "row", axes = 1, top = 10)
# Contribuciones de filas a la dimension 2
fviz_contrib(res.ca, choice = "row", axes = 2, top = 10)
# Contribucion total
fviz_contrib(res.ca, choice = "row", axes = 1:2, top = 10)

fviz_ca_row(res.ca, col.row = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)

# Analisis de columnas
col <- get_ca_col(res.ca)
col

# coordenadas
head(col$coord)
fviz_ca_col(res.ca, repel = TRUE)
# columnas con perfiles similares son agrupadas
# Mayor distancia del origen implica mejor representacipn


# Cos2: Asociacion de columnas con dimensiones
head(col$cos2)
fviz_ca_col(res.ca, col.col = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
library("corrplot")
corrplot(col$cos2, is.corr=FALSE)
fviz_cos2(res.ca, choice = "col", axes = 1:2)


```

```r
# Contribuciones a las dimensiones
head(col$contrib)

# para explicar la variabilidad del conjunto de datos
library("corrplot")
corrplot(col$contrib, is.corr=FALSE)
# Contribuciones de columnas a la dimension 1
fviz_contrib(res.ca, choice = "col", axes = 1, top = 10)
# Contribuciones de columnas a la dimension 2
fviz_contrib(res.ca, choice = "col", axes = 2, top = 10)
# Contribucion total
fviz_contrib(res.ca, choice = "col", axes = 1:2, top = 10)

fviz_ca_col(res.ca, col.col = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)

## Biplots

# Simetrico
fviz_ca_biplot(res.ca, repel = TRUE,ylim=c(-0.4,0.4))

# Asimetrico
fviz_ca_biplot(res.ca,
map ="rowprincipal", arrow = c(TRUE, TRUE),
repel = TRUE,xlim=c(-0.4,0.4),ylim=c(-0.4,0.4))
summary(res.ca)
help(fviz_ca_biplot)

# Biplot de contribuciones
fviz_ca_biplot(res.ca, map ="colgreen", arrow = c(TRUE, FALSE),
repel = TRUE)
```