Practico 3

Ejercicio 2

A) Aplicar al menos un método jerárquico, por ejemplo: vecino mas cercano, vecino mas lejano, vinculación promedio o Ward.

i) Describir muy brevemente en qué consiste.

El análisis de conglomerados (clusters) es un metodo que tiene por objeto agrupar elementos homogéneos en función de las similitudes o similaridades entre ellos teniendo en cuenta un conjunto de variables especificas. Normalmente se agrupan las observaciones, pero el análisis de conglomerados puede también aplicarse a las variables. Permite clasificar las observaciones en grupos que son homogeneos internamente y heterogeneos entre ellos.

El metodo jerarquico tiene por objetivo estructurar los elementos de un conjunto ordenados en funcion de su similitud. Esto implica que los datos se ordenan en niveles, de manera que los superiores contienen a los inferiores. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos, sin embargo, la jerarquía construida permite obtener una partición de los datos en grupos.

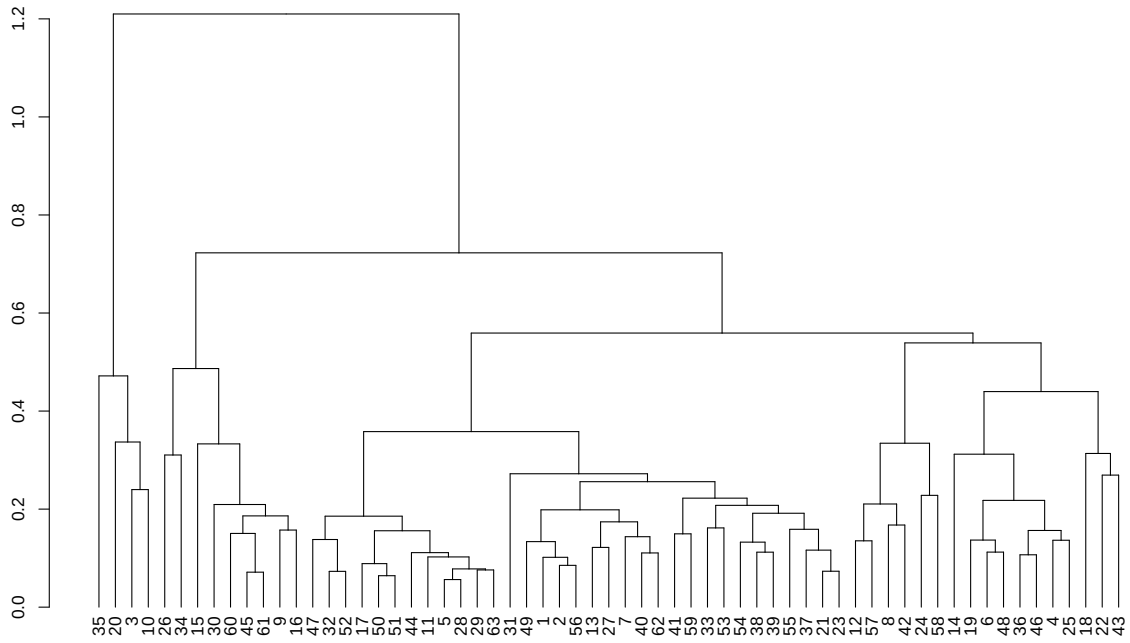
Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en una distancia.

Para estos datos se ajustaron clusters empleando los metodos de vecino mas cercano, vecino mas lejano y vinculacion promedio. Hay que evaluar hasta qué punto la estructura del dendrograma creado refleja las distancias originales entre observaciones. Una forma de hacerlo es empleando el coeficiente de correlación entre las distancias copheneticas del dendrograma (altura de los nodos) y la matriz de distancias original. Cuanto más cercano es el valor a 1, mejor refleja el dendrograma la verdadera similitud entre las observaciones. Valores superiores a 0.75 suelen considerarse como buenos. Esta medida puede emplearse como criterio de ayuda para escoger entre los distintos métodos de linkage. En R, la función cophenetic() calcula las distancias copheneticas de un hierarchical clustering.

Para estos datos, el método de linkage average o vinculacion promedio consigue representar ligeramente mejor la similitud entre las observaciones, el coeficiente de correlacion para este metodo es de 0,81 superando el de vecino mas cercano (0,79) y el de vecino mas lejano (0,75).

Este metodo obtiene la distancia entre dos grupos calculando la distancia promedio entre todos los pares de observaciones que pueden formarse tomando un miembro de un grupo y otro miembro del otro grupo. Asi, va formando los grupos y calculando sucesivamente la matriz de distancias hasta que todas las observaciones quedan en un solo grupo. A continuacion se grafica el dendrograma obtenido:

ii) Graficar el dendograma.

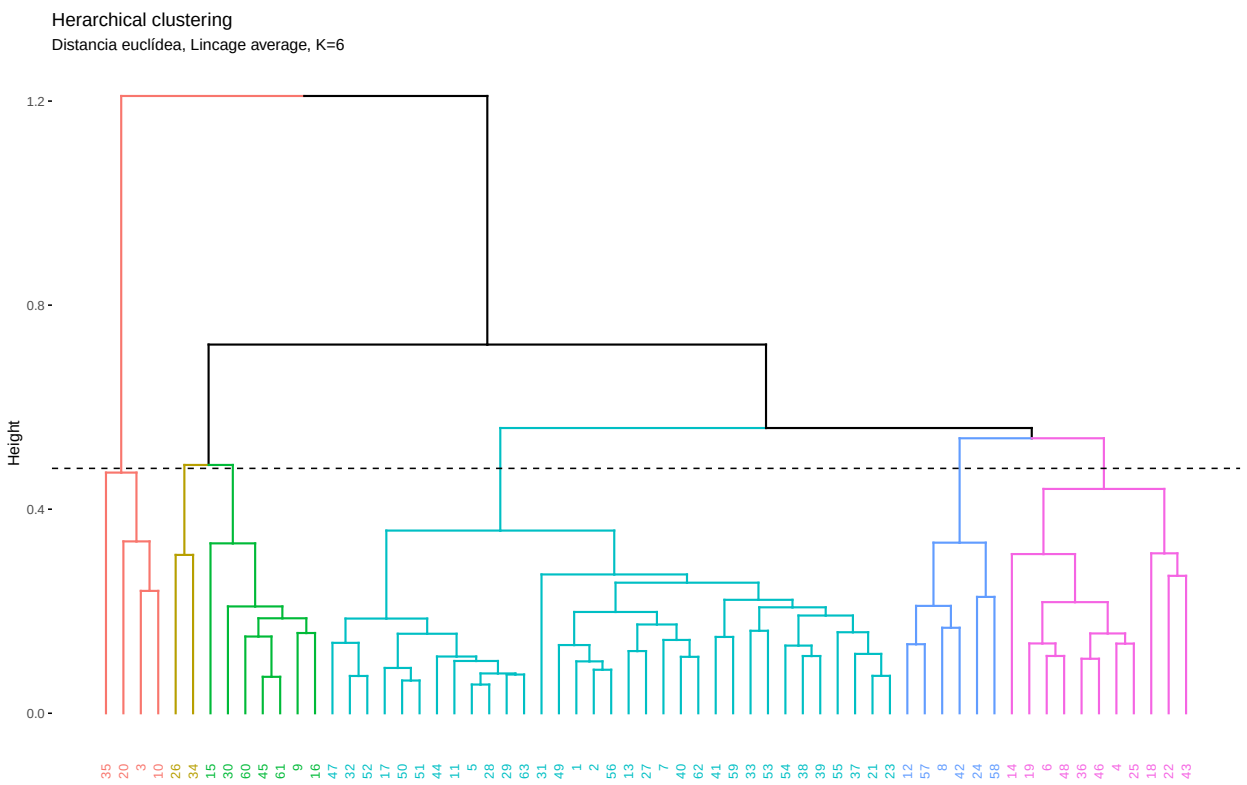


En el eje horizontal del dendrograma tenemos cada uno de los datos que componen el conjunto de entrada, mientras que en el eje vertical se representa la distancia euclídea que existe entre cada grupo a medida que éstos se van jerarquizando. Cada línea vertical del diagrama representa un agrupamiento que coincide con los puntos arropados por ésta, y como se ve en el dendrograma, estos van formándose progresivamente hasta tener un solo gran grupo determinado por la línea horizontal superior. Así, al ir descendiendo en la jerarquía, vemos que de un solo grupo pasamos a 2, luego a 3, luego a 6, y así sucesivamente.

iii) Determinar cuántos grupos se seleccionan. Justificar, por ejemplo usando el Criterio de Hartigan o el Coeficiente de Correlación Cofenético.

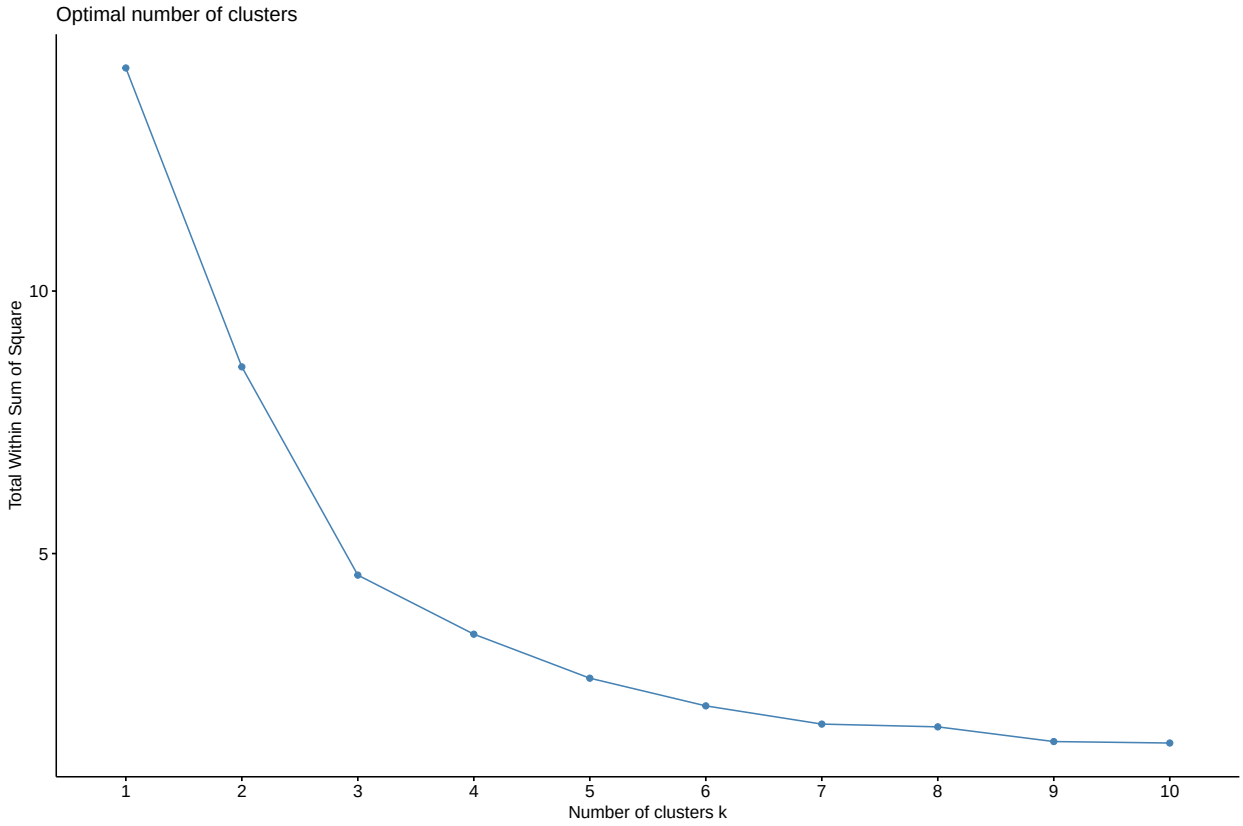
Para determinar el numero de clusters, se opto por considerar la diferencia entre la Suma de Cuadrados Dentro con g y g+1 grupos, analizando la reduccion de la variabilidad relativa tras un agrupamiento adicional. Nos guiamos por el criterio de Hartigan, quien sugiere introducir un grupo mas si $F > 10$.

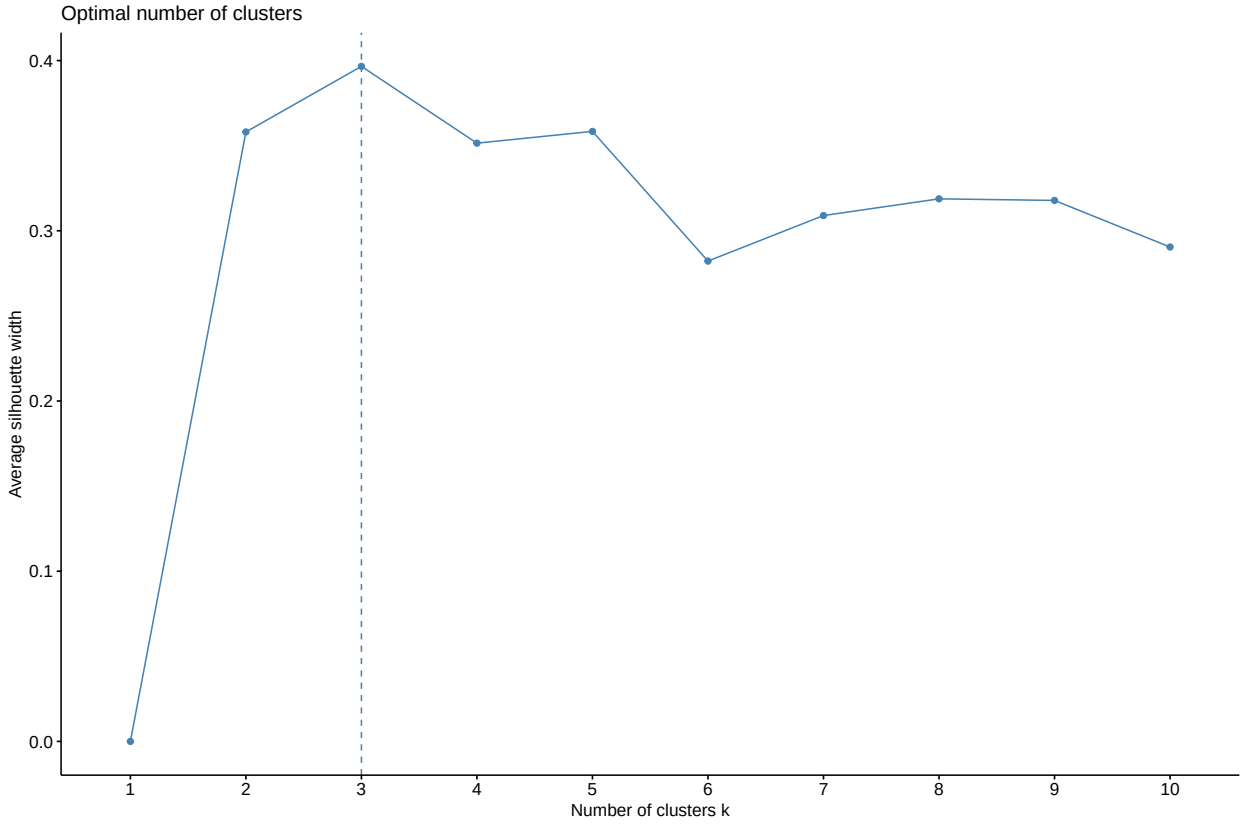
Aplicando el criterio de Hartigan nos quedamos con 6 grupos. La siguiente imagen muestra el corte a la altura de 0,48, con lo que se obtienen los 6 clusters.



B) Aplicar el método k-means

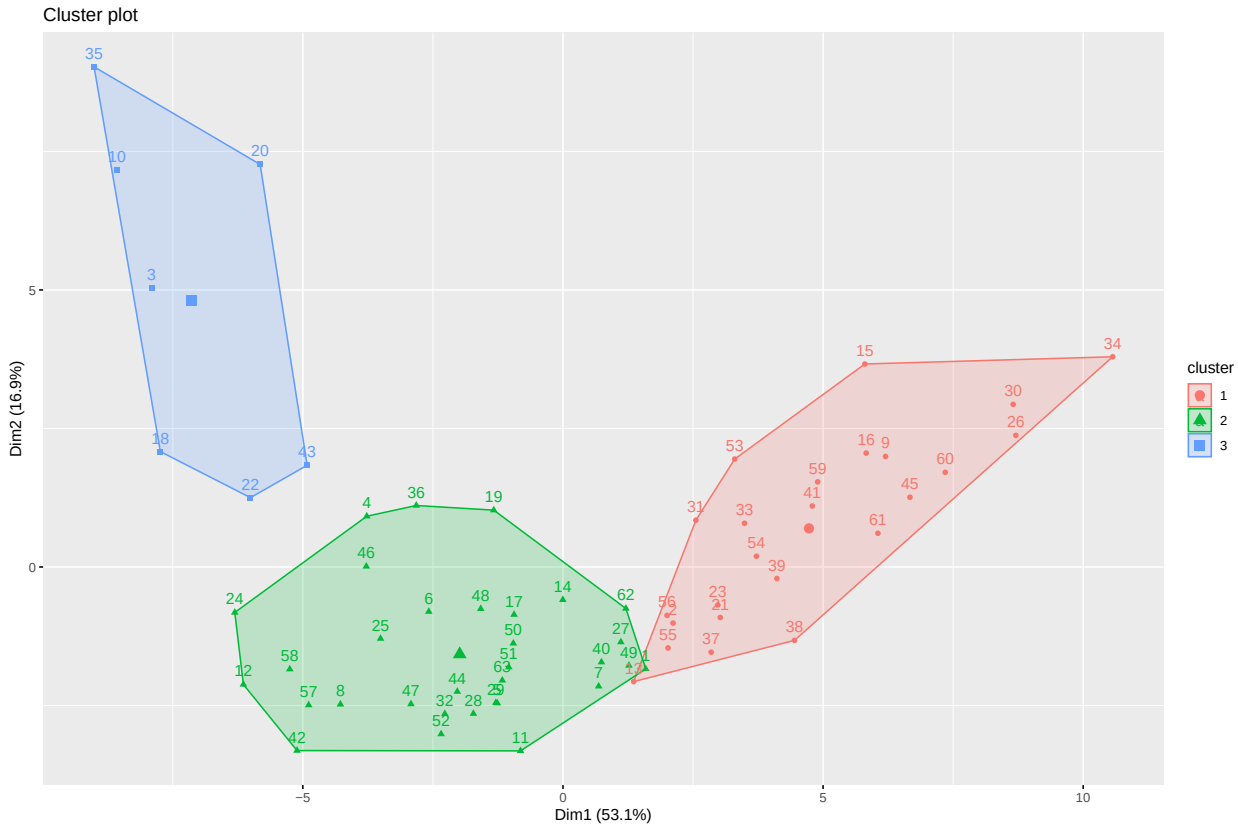
Para aplicar el metodo de k means, procedemos a determinar el numero optimo de clusters, esto se llevo a cabo a por medio del metodo del codo y de silhouette.





A partir de los graficos decidimos optar por un k de 3 para este metodo. Aplicando el metodo para el k

obtenido los resultados son:



i) Indicar cómo se puede atemperar la influencia de la elección de los centroides iniciales en los resultados.

Para atemperar la influencia de la eleccion de los centroides iniciales en los resultados, se puede optar por el algoritmo de K medias recortado.

ii) Determinar si se encontró una estructura de agrupación fuerte o débil; por ejemplo, empleando el Coeficiente de Silhouette.

El coeficiente de silhouette para un k igual a 3 (optimo), arrojo un valor de 0.39, lo cual indica una estructura debil y podria ser artificial.

C) En base a alguno de los agrupamientos previos, realizar un análisis descriptivo comparativo entre grupos.

La comparacion se realiza a partir del metodo de k means.

i) En base al mismo, caracterice a los grupos.

```
## K-means clustering with 3 clusters of sizes 24, 32, 7
##
## Cluster means:
```

```
##      inicial    priminc   primcomp     secinc   seccomp     terinc    tercomp
## 1 0.04565302 0.16967928 0.31747032 0.23747524 0.1371863 0.01649390 0.02312523
## 2 0.02914255 0.07598793 0.18024341 0.18045492 0.1903443 0.03698935 0.06985867
## 3 0.02027174 0.02338237 0.06720558 0.07752272 0.1624247 0.04800302 0.07988275
##      univinc   univcomp       posinc     poscomp  sabeleer nosabeleer cobertura
## 1 0.03371905 0.01642014 0.0008485351 0.001929029 0.9606856 0.03931435 0.4849189
## 2 0.10736246 0.10822686 0.0061651023 0.015224436 0.9802584 0.01974164 0.7513673
## 3 0.29435873 0.18124906 0.0176795450 0.028019751 0.9892398 0.01076023 0.8174329
##   nocobertura  ocupados desocupados inactivos       nbi1        nbi2
## 1   0.5150811 0.5713418  0.05757306 0.3710852 0.16063104 0.019346430
## 2   0.2486327 0.5873205  0.03959662 0.3730829 0.04530515 0.007893113
## 3   0.1825671 0.5948472  0.03691182 0.3682410 0.02222140 0.031866630
##          nbi3         nbi4        nbi5   sinpriv    privpat    privrec
## 1 0.019653649 0.0009969910 0.011301168 0.7558555 0.18051871 0.03648780
## 2 0.003701317 0.0005312054 0.002357384 0.9132672 0.05036912 0.02362635
## 3 0.002164994 0.0003078090 0.001157269 0.8306187 0.04217082 0.10105129
##      privconv      casa      rancho     casilla departamento piezaolocal
## 1 0.025913952 0.9016102 0.006490445 0.010785148   0.07100501  0.01010923
## 2 0.005663855 0.8316672 0.002237083 0.002434295   0.15695366  0.00670777
## 3 0.007039446 0.2718321 0.001670866 0.001694924   0.68957793  0.03522420
##         uno       dos      tres     cuatro      cinco       seis       siete
## 1 0.1343976 0.1574170 0.1873381 0.19661102 0.14336784 0.08502435 0.061731458
## 2 0.1563537 0.2365881 0.2066777 0.19107655 0.11479263 0.05262294 0.029721598
## 3 0.3454740 0.3351204 0.1542340 0.09322318 0.04517867 0.01566020 0.007891112
##      ochoomas
## 1 0.034112589
## 2 0.012166838
## 3 0.003218411
##
## Clustering vector:
##  [1] 2 1 3 2 2 2 2 2 1 3 2 2 1 2 1 1 2 3 2 3 1 3 1 2 2 1 2 2 2 1 1 2 1 1 3 2 1 1
## [39] 1 2 1 2 3 2 1 2 2 2 2 2 2 2 1 1 1 1 2 2 1 1 1 2 2
##
## Within cluster sum of squares by cluster:
## [1] 1.7226313 2.0679164 0.7990961
##  (between_SS / total_SS =  67.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

A partir de esta salida podemos ver la media para los proporciones de cada una de las variables en cada uno de los grupos y de esta forma ver cuales son las variables que mas influyen para separar a los grupos.

Por otra parte tambien se puede apreciar el ratio de la suma de cuadrados entre-clusters y la suma de cuadrados totales. Este ratio es equivalente al R2 de los modelos de regresión e indica el porcentaje de varianza explicada por el modelo respecto al total de varianza observada. Puede utilizarse para evaluar el clustering obtenido pero, al igual que ocurre con R2 al incrementar el número de predictores, el ratio between_SS / total_SS aumenta con el número de clusters creados. Esto último se debe de tener en cuenta para evitar problemas de overfitting. En este caso el porcentaja de la varianza explicada por el modelo respecto al total de la varianza observada es de 67,8%.

Interpretacion de las proporciones medias para los distintos grupos: Respecto al nivel educativo podemos ver el grupo 1 posee mayor proporcion de personas con primario completo y secundario incompleto, por lo que seria una buena indicadora de las personas que han alcanzado este nivel educativo. Por otra parte el grupo 3 posee una mayor proporcion de personas en los grupos que se encuentran relacionados a la universidad completa e incompleta, obteniendo un porcentaje mayoritario en universidad incompleta, por lo que seria un grupo representado por personas que han alcanzado formarse en una carrera de grado. El grupo 2 posee una proporcion mayoritaria en los grupos que han alcanzado la universidad al igual que el grupo 3, pero con valores menores.

Para la variable que refleja si las personas saben leer o no saben leer, podemos ver que ninguno de los grupos permite una buena division de estas categorias y los valores en todos los grupos son homogeneos.

En la cobertura de salud podemos ver que el grupo 1 no permite realizar una buena division para esta variable, mientras que los grupos 2 y 3 poseen grandes porcentajes de personas con cobertura.

En todos los grupos el porcentaje de ocupados, desocupados e inactivos es muy similar.

Para las necesidades basicas el unico grupo que se caracteriza por concentrar los nb1 es el grupo 1, las demas categorias tienen valores similares. Recordemos que esta categoria no suma 1 porque la mayor proporcion de las personas posee sus necesidades basicas satisfechas.

En todos los grupos la privacion del material en el hogar es similar y encuentra mayores proporciones en sin privacion, se destaca una concentracion de personas privpat en el grupo 1 y con privrec en el grupo 2.

El grupo 1 y 2 poseen la mayor proporcion de personas que viven en casas, mientras que el grupo 3 relfeja las personas que viven en departamentos.

La cantidad de habitantes en los hogares es similar para todos los grupos, sin embargo, se destaca en el grupo 3 que la mayor proporcion la obtienen las personas que viven solas y en pareja, mientras que para los grupos 1 y 2 todos los valores son muy similares.

ii) ¿Se le pueden dar "nombres" intuitivos a los grupos?

Grupo 1: Concentra personas que estuvieron por comenzar el secundario o bien que se encuentran en un proceso de aprendizaje secundario o que nunca pudieron terminarlo y que viven en casas.

Grupo 2: Concentra personas que han alcanzado la formacion de grado (universitaria), ya sea si la han temrinado o no, que viven en casas y que poseen cobertura de salud.

Grupo 3: Concentra personas que han alcanzado la formacion de grado (universitaria), ya sea si la han terminado o no, que viven solos o en pareja en departamentos y que poseen cobertura medica.

De esta forma se podria dar nombres intuitivos a los grupos focalizandonos en la educacion y en la descripcion previamente realizada:

Grupo 1: ciudadanos cordobeces con trabajos de oficio. Asumo que son cordobeses como consecuencia de que residen en casas, dado a que se encuentran inmersos en no mas que la educacion secundaria asumo que la casa fue heredada o bien siempre vivieron alli por lo que son nacidos en la provincia y que realizan trabajos de oficio como consecuencia de no poseer educacion universitaria. Algo que fortalece mi hipotesis es la cobertura de salud, a partir de la cual podemos observar que la mitad de las personas de este grupo no se encuentran cubiertas, lo que se justificaria en el caso de trabajos informales o en negro y ademas son el unico grupo con una proporcion mayoritaria de necesidades basicas nbi1 insatisfechas, lo que refleja su situacion eceonomica.

Grupo 2: ciudadanos adultos con trabajos profesionales. Asumo que son ciudadanos adultos dado a que han podido adquirir una casa, la cual justifico por medio de que la persona ha podido trabajar un tiempo prologando gracias a que posee un titulo de grado que le permitio ahorrar para acceder a su hogar propio y se encuentra en proceso de formar una familia (como consecuencia de que la menor proporcion de habitantes en el hogar es la de uno solo en comparacion con dos, tres y cuatro).

Grupo 3: ciudadanos jovenes con trabajos profesionales. Asumo que es un ciudadano joven como consecuencia de que habita un departamento, este puede ser de cordoba o no, como consecuencia de que esta es una provincia conocida por recibir muchos estudiantes que vienen de otras provincias a formarse universitariamente y al finalizar sus estudios normalmente se quedan residiendo en la provincia. Por lo general estos jovenes viven en departamentos y dado a la posibilidad que tubieron de estudio univesitario asumo que poseen un trabajo de tipo profesional.

D) Indicar si se encuentra alguna conexión entre los resultados obtenidos en los incisos anteriores y los obtenidos en el Problema 2 de la guía II.

Encuentro una notable relacion con los resultados de la guia 2, en esta guia el mejor analisis de componentes pricipales resumia el 80% de la variabilidad de los datos con 3 factores y en estos tres factores la variables de mayor relevancia era en cada caso: * Componente 1: nivel educativo. * Componente 2: tipo de hogar y cantidad de habitantes. * Componente 3: necesidades basicas insatisfechas. A partir de esto podemos ver que las variables de mayor relevancia en el analisis de componentes principales son las variables que mayor importancia tienen a la hora de formar clusters.

i) Opcional: realizar un análisis factorial exploratorio breve usando esta base reducida para que sean más comparables los resultados.

```
##
## Call:
## factanal(x = datos2, factors = 2, data = datos2, scores = "Bartlett",     rotation = "none")
##
## Uniquenesses:
##      inicial      priminc     primcomp       secinc      seccomp       terinc
##        0.804        0.098        0.105        0.311        0.547        0.179
##      tercomp      univinc     univcomp       posinc     sabeleer    cobertura
##        0.175        0.157        0.366        0.308        0.089        0.013
##     ocupados  desocupados         nbi1         nbi2         nbi3         nbi4
##        0.893        0.229        0.060        0.681        0.265        0.838
##         nbi5      sinpriv      privpat      privrec         casa       rancho
##        0.273        0.330        0.348        0.469        0.005        0.687
##      casilla departamento          uno          dos         tres       cuatro
##        0.822        0.005        0.239        0.118        0.390        0.189
##        cinco         seis        siete
##        0.195        0.116        0.143
##
## Loadings:
##              Factor1 Factor2
## inicial       0.232   0.377
## priminc       0.700   0.642
## primcomp      0.770   0.550
## secinc        0.757   0.341
## seccomp      -0.178  -0.649
## terinc       -0.722  -0.547
## tercomp      -0.557  -0.718
## univinc      -0.901  -0.178
## univcomp     -0.560  -0.566
## posinc       -0.782  -0.283
## sabeleer     -0.654  -0.696
## cobertura    -0.613  -0.782
```

```
## ocupados     -0.217  -0.243
## desocupados   0.459   0.749
## nbi1          0.595   0.766
## nbi2         -0.266   0.498
## nbi3          0.498   0.698
## nbi4          0.283   0.286
## nbi5          0.479   0.706
## sinpriv      -0.157  -0.803
## privpat       0.399   0.702
## privrec      -0.620   0.383
## casa          0.973  -0.220
## rancho        0.288   0.480
## casilla       0.184   0.380
## departamento -0.983   0.172
## uno          -0.858   0.161
## dos          -0.870  -0.353
## tres          0.506  -0.595
## cuatro        0.855  -0.282
## cinco         0.896
## seis          0.854   0.392
## siete         0.740   0.556
##
##                Factor1 Factor2
## SS loadings     13.512   9.044
## Proportion Var   0.409   0.274
## Cumulative Var   0.409   0.684
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 2203.07 on 463 degrees of freedom.
## The p-value is 6.6e-224
```

Podemos ver que para este analisis factorial con 2 factores, se cumple practicamente lo mismo que habiamos establecido para el analisis de componentes principales, vemos que para los dos factores que resumen un total del 68,4% de la variabilidad total de los datos las variables mas importantes son: nivel educativo, hogar y cantidad de personas, que son las mismas que permiten realizar la division de los clusters.