Parcial III

Alimi Garmendia 14-10392

3/13/2020

Tarea : Ejercicio 3

Exámen: Ejercicios 1,2,4,6,7

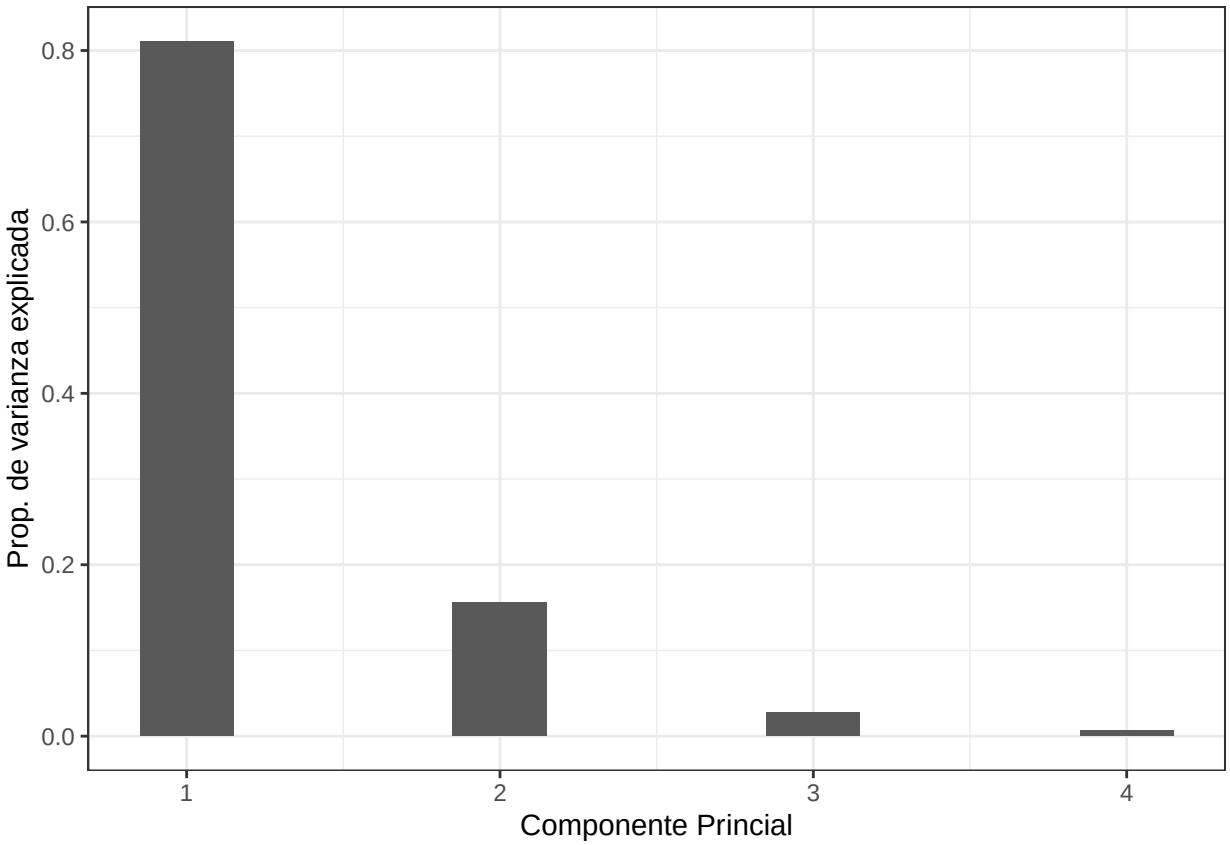
1. Análisis de Componentes Principales: Realice un análisis de componentes principales sobre los porcentajes de votación por partido. Este debe incluir con cuantos componentes se debe trabajar luego del análisis y que expresan cada uno de esos componentes.

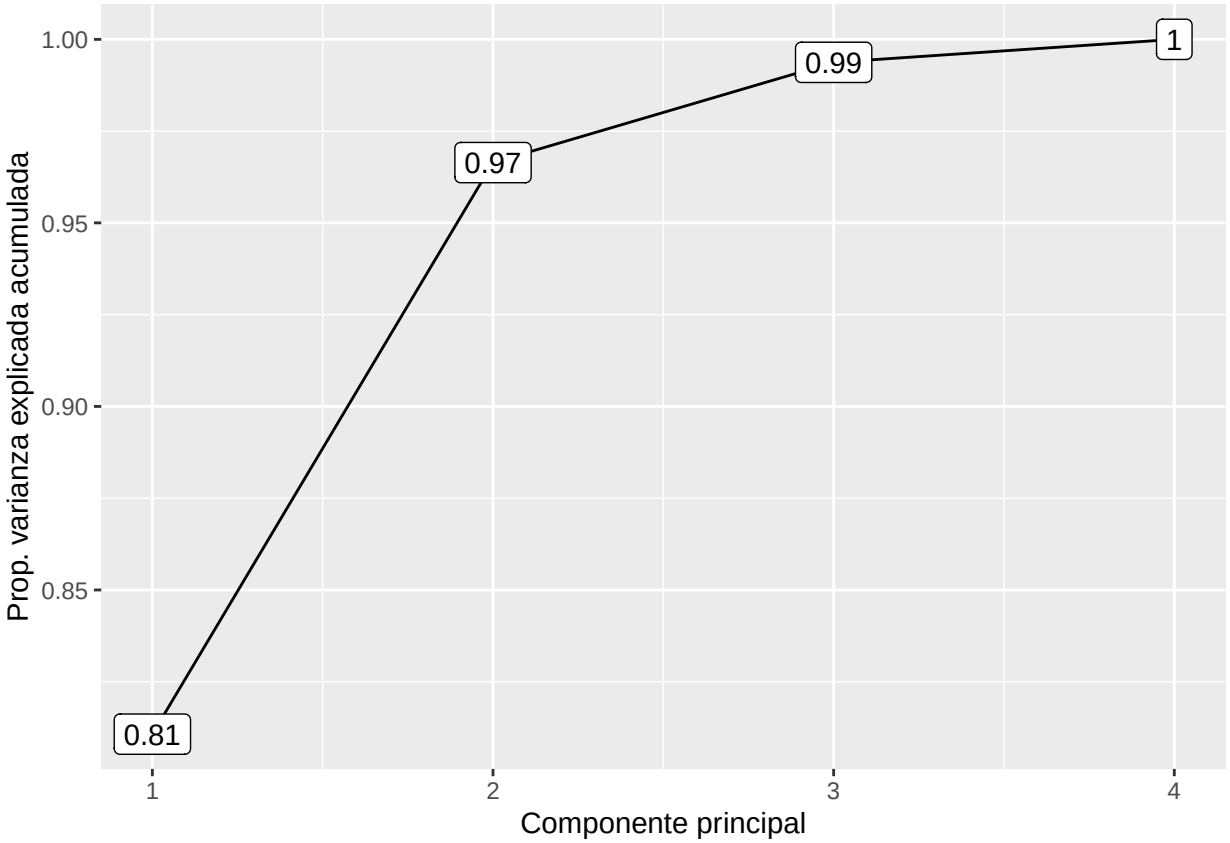
Comenzamos análizando los datos para el pártido demócrata.

```
demo <- data %>%
  select(Estado,Cod,DEM12,DEM04,DEM92,DEM80,Tend,REG)

pca.demo <- princomp(demo[,3:6],cor = T)
summary(pca.demo)
```

```
## Importance of components:
##                          Comp.1 Comp.2 Comp.3   Comp.4
## Standard deviation     1.8008 0.7892 0.3305 0.158345
## Proportion of Variance 0.8107 0.1557 0.0273 0.006268
## Cumulative Proportion  0.8107 0.9664 0.9937 1.000000
```

Como podemos ver con sólo dos componentes podemos representar aproximadamente 96% de la varianza de los datos Podemos verlo de manera gráfica con las siguientes figuras


0.8
0.6
0.4
0.2
0.0
Prop. de varianza explicada
1
2
3
4
Componente Princial



Como era de esperarse, la primera componente es la que describe la mayor parte de la varianza. De igual forma podemos ver que con dos componentes podemos describir una gran porción de la varianza. Si revisamos el criterio de Keiser.

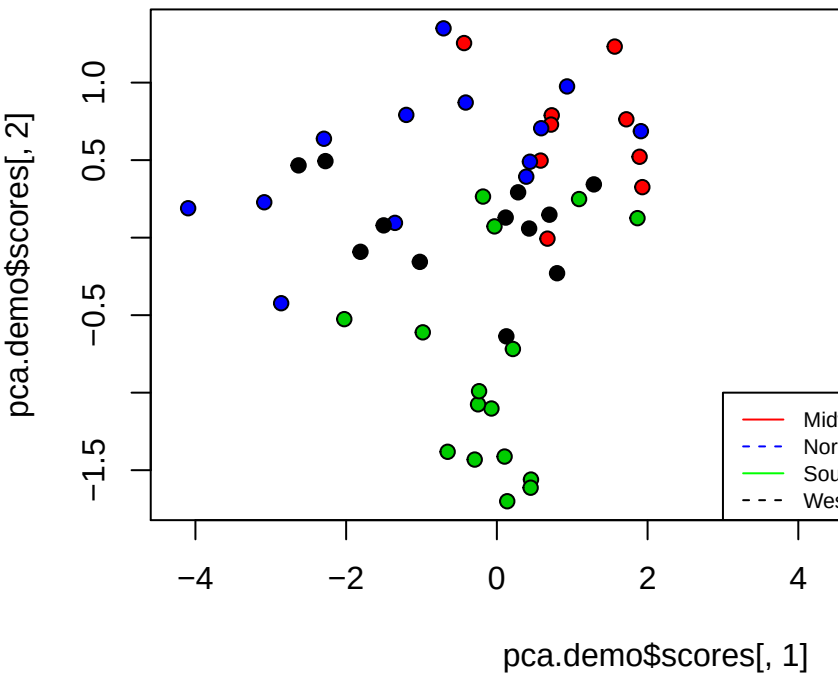
```
eigen(cor(demo[,3:6]))$values
```

```
## [1] 3.24288 0.62283 0.10921 0.02507
```

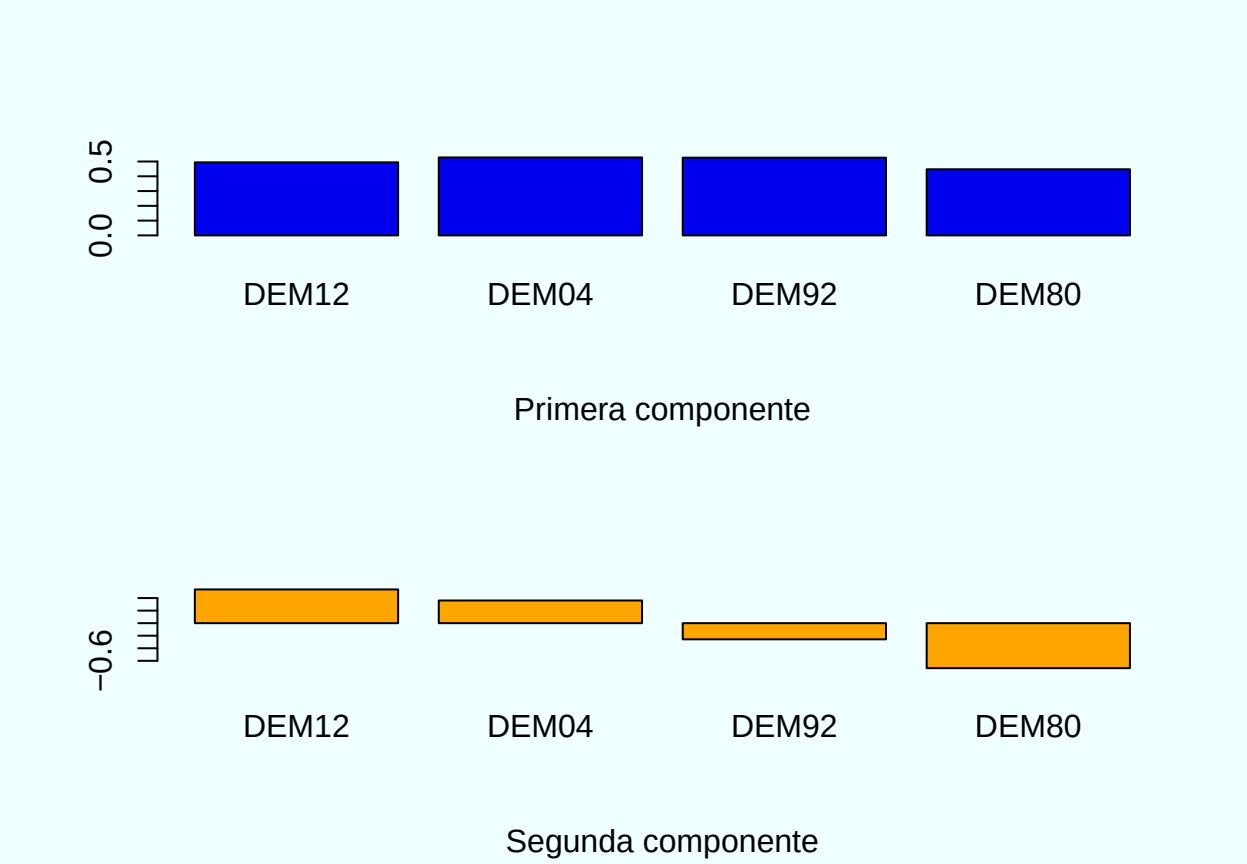
```
mean(eigen(cor(demo[,3:6]))$values)
```

```
## [1] 1
```

A pesar de ser menor que la media, la segunda componente es tomada pues añade un importante 16% a la varianza acumulada, sin embargo las demás componentes pueden ser obviadas.



Podemos visualizar las dos componentes de la forma:



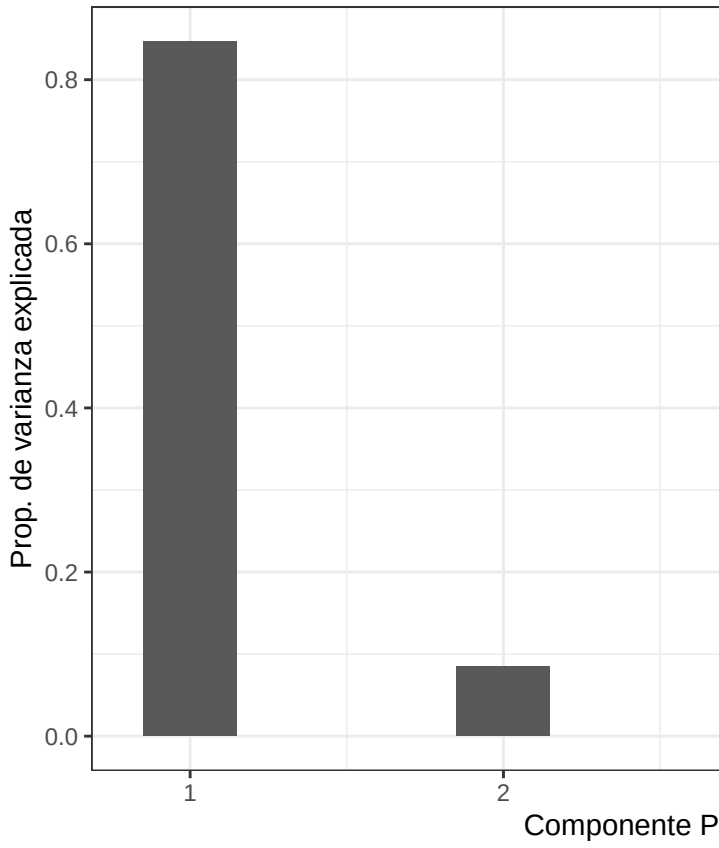
Como hemos visto, con dos componentes es suficiente para representar nuestros datos

Ahora realizamos el análisis para el pártido repúblicano

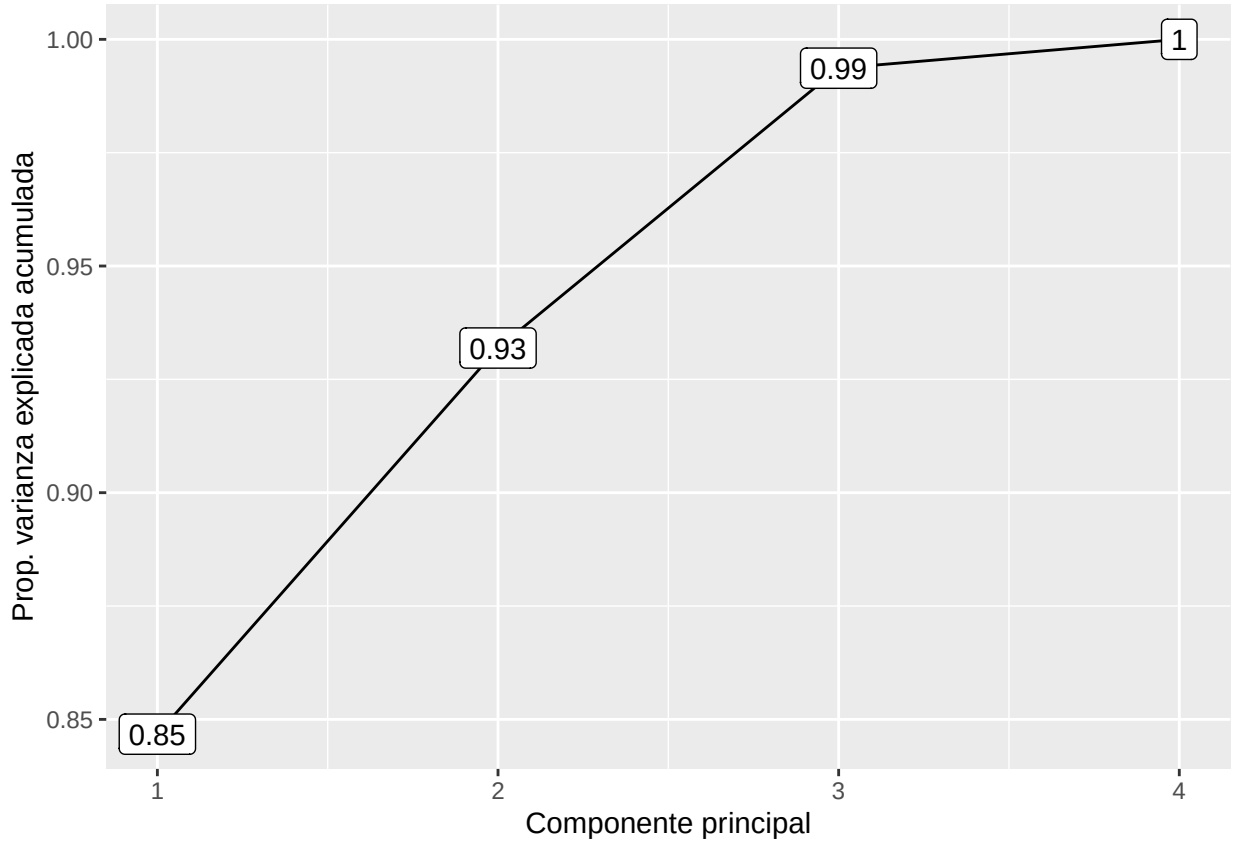
```
gop <- data %>%
  select(Estado,Cod,GOP12,GOP04,GOP92,GOP80,Tend,REG)

pca.gop <- princomp(gop[,3:6],cor = T)
summary(pca.gop)
```

```
## Importance of components:
##                          Comp.1  Comp.2  Comp.3   Comp.4
## Standard deviation     1.8404 0.58369 0.49690 0.159531
## Proportion of Variance 0.8467 0.08517 0.06173 0.006363
## Cumulative Proportion  0.8467 0.93191 0.99364 1.000000
```

Como podemos ver con sólo dos componentes podemos representar aproximadamente 93% de la varianza de los



datos Podemos verlo de manera gráfica con las siguientes figuras



Como era de esperarse, la primera componente es la que describe la mayor parte de la varianza. De igual forma podemos ver que con dos componentes podemos describir una gran porción de la varianza. Si revisamos el criterio de Keiser.

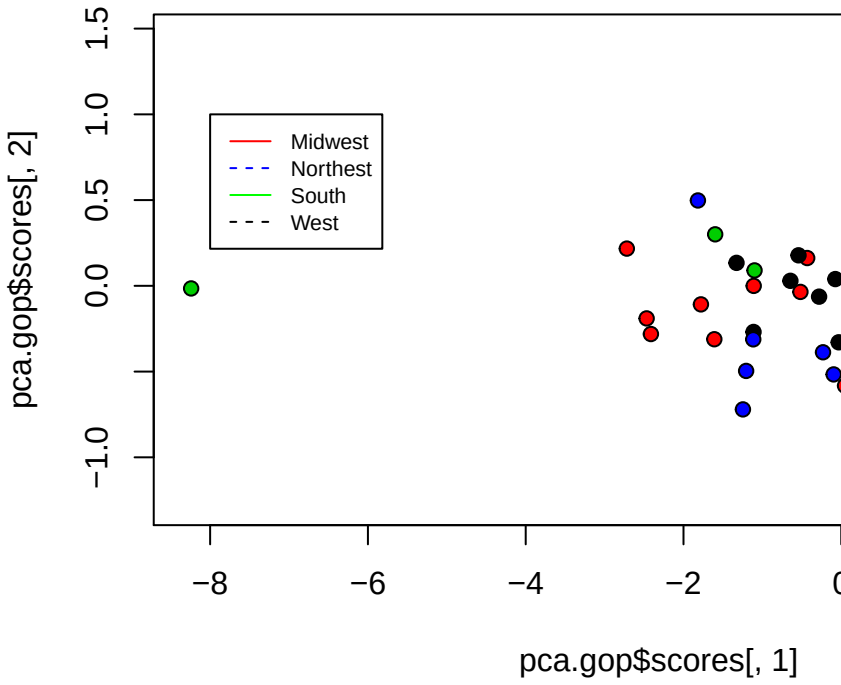
```
eigen(cor(gop[,3:6]))$values
```

```
## [1] 3.38695 0.34069 0.24691 0.02545
```

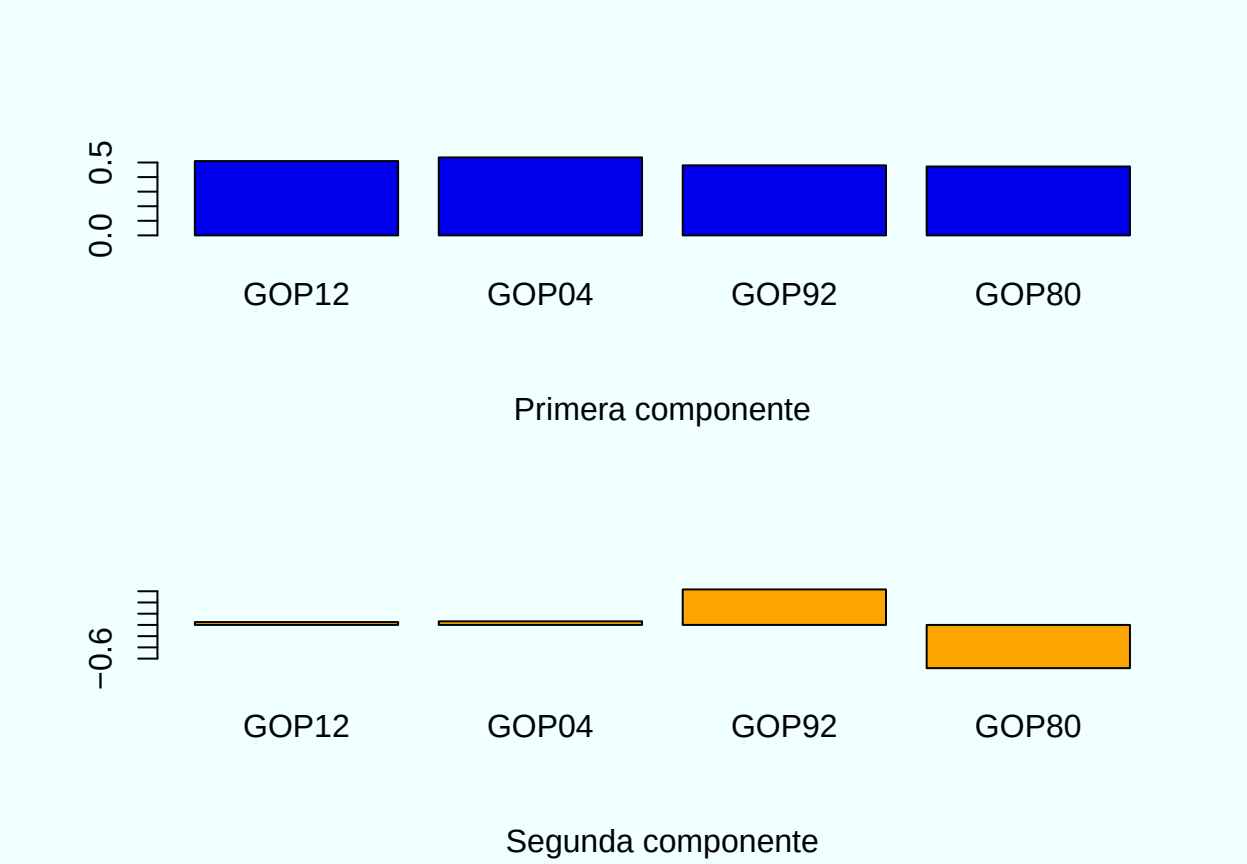
```
mean(eigen(cor(gop[,3:6]))$values)
```

```
## [1] 1
```

A pesar de ser menor que la media, la segunda componente es tomada pues añade un aceptable 8% a la varianza acumulada, sin embargo las demás componentes pueden ser obviadas.



Podemos visualizar las dos componentes de la forma:



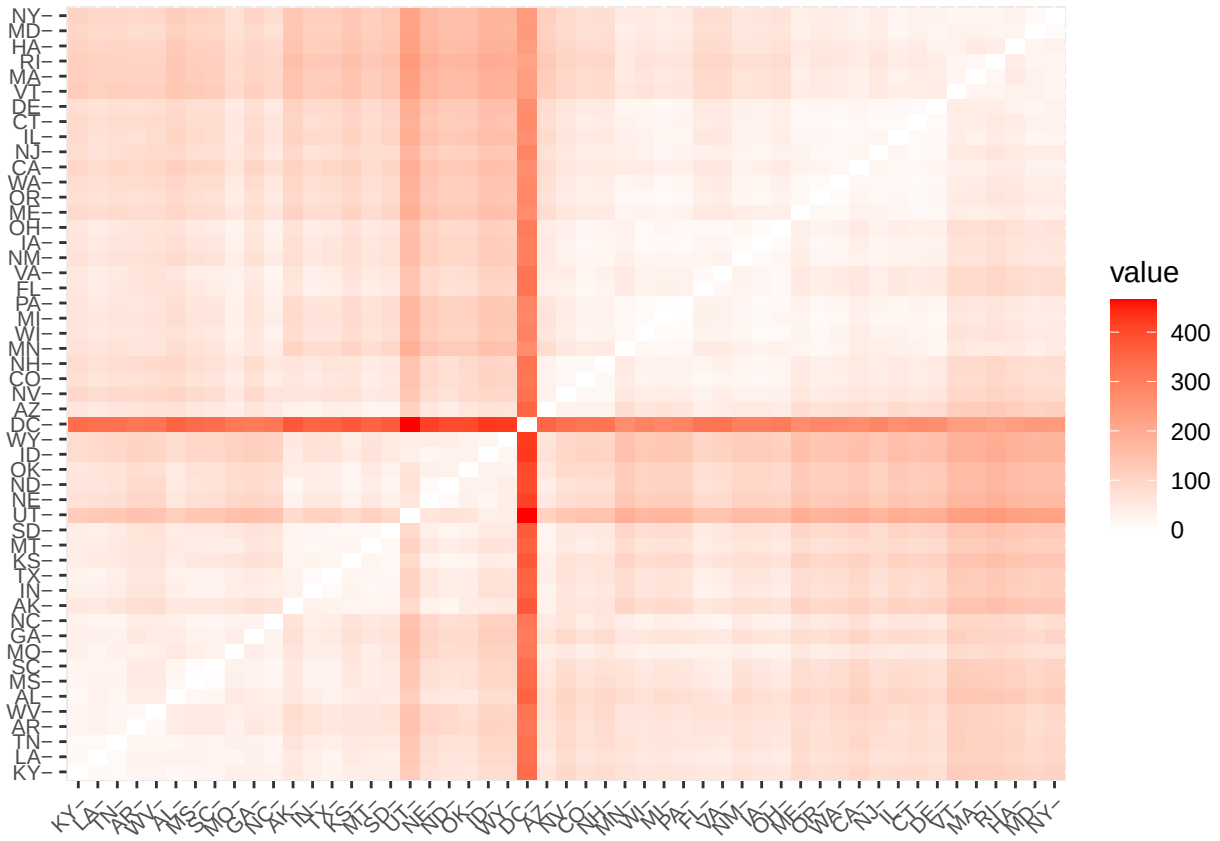
2. Análisis de Conglomerados: Realice un análisis de conglomerados utilizando dos métodos jerárquicos + el método de k-medias para agrupar los 50 estados + el Distrito de Columbia según los porcentajes de votación. Exprese el número de conglomerados que escogería, y que entidades estarían en cada uno de esos conglomerados.

Para el cálculo de las distancias en ambos métodos utilizaremos la distancia de Manhattan.Comenzaremos usando el método de K-means para varios K's

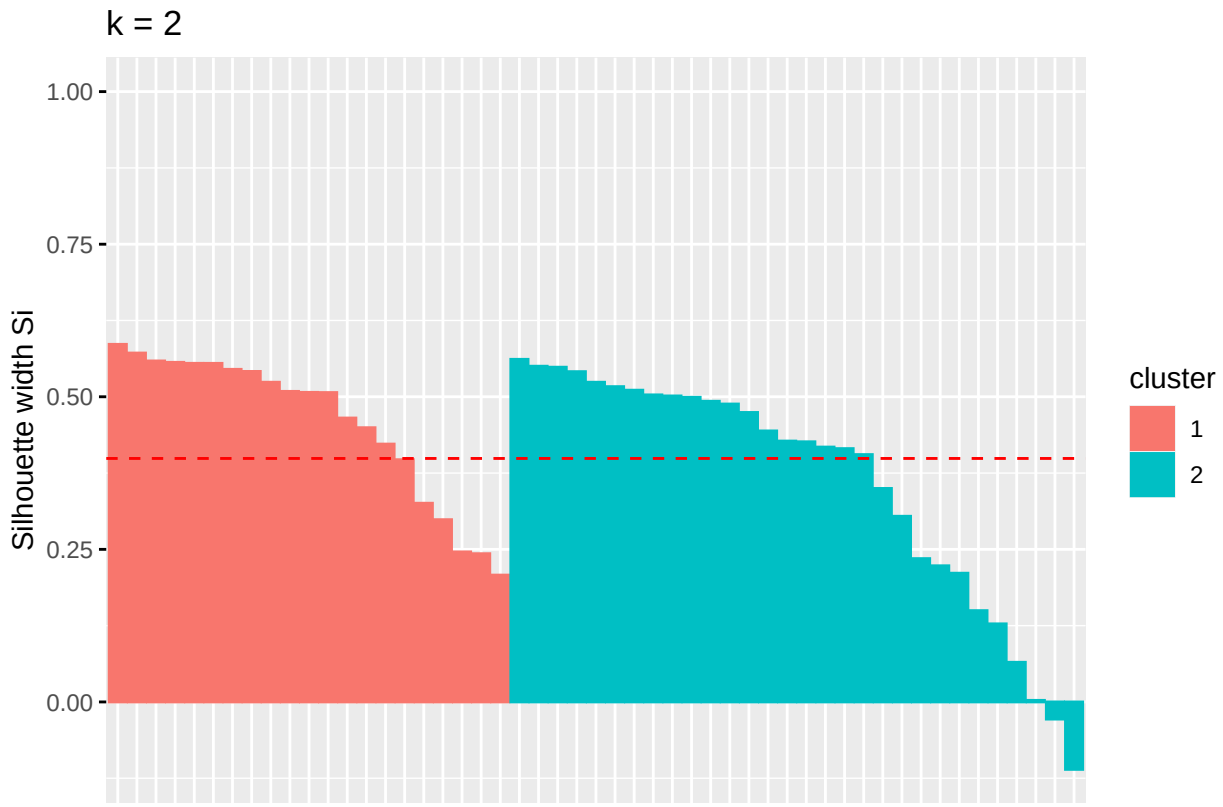
```
##            Estado Cod DEM12 GOP12 DEM04 GOP04 DEM92 GOP92 DEM80 GOP80 Tend
## 1         Alabama  AL 38.36 60.55 36.84 62.46 40.88 47.65 47.45 48.75  REP
## 2          Alaska  AK 40.81 54.80 35.52 61.07 30.29 39.46 26.41 54.35  REP
## 3         Arizona  AZ 44.59 53.65 44.40 54.87 36.52 38.47 28.24 60.61  REP
## 4        Arkansas  AR 36.88 60.57 44.55 54.31 53.21 35.48 47.52 48.13  REP
## 5      California  CA 60.24 37.12 54.31 44.36 46.01 32.61 35.91 52.69  DEM
## 6        Colorado  CO 51.49 46.13 47.02 51.69 40.13 35.87 31.07 55.07  BGS
## 7     Connecticut  CT 58.06 40.73 54.31 43.95 42.21 35.78 38.52 48.16  DEM
## 8        Delaware  DE 58.61 39.98 53.35 45.75 43.52 35.78 44.87 47.21  DEM
## 9            D.C.  DC 90.91  7.28 89.18  9.34 84.64  9.10 74.89 13.41  DEM
## 10        Florida  FL 50.01 49.13 47.09 52.10 39.00 40.89 38.50 55.52  BGS
## 11        Georgia  GA 45.58 53.30 41.37 57.97 43.47 42.88 55.76 40.95  REP
## 12         Hawaii  HA 70.55 27.84 54.01 45.26 48.09 36.70 44.80 42.90  DEM
## 13          Idaho  ID 32.62 64.53 30.26 68.38 28.42 42.03 25.19 66.46  REP
## 14       Illinois  IL 57.60 40.73 54.82 44.48 48.58 34.34 41.72 49.65  DEM
## 15        Indiana  IN 43.93 54.13 39.26 59.94 36.79 42.91 37.65 56.01  REP
## 16           Iowa  IA 51.99 46.18 49.23 49.90 43.29 37.27 38.60 51.31  BGS
```

```
## 17         Kansas  KS 37.99 59.71 36.62 62.00 33.74 38.88 33.29 57.85  REP
## 18       Kentucky  KY 37.80 59.71 39.69 59.55 44.55 41.34 47.61 49.07  REP
## 19      Louisiana  LA 40.58 57.78 42.22 56.72 45.58 40.97 45.75 51.20  REP
## 20          Maine  ME 56.27 40.98 53.57 44.58 38.77 30.39 42.25 45.61  DEM
## 21       Maryland  MD 61.97 35.90 55.91 42.93 49.80 35.62 47.14 44.18  DEM
## 22  Massachusetts  MA 60.65 37.51 61.94 36.78 47.54 29.03 41.75 41.90  DEM
## 23       Michigan  MI 54.21 44.71 51.23 47.81 43.77 36.38 42.50 48.99  BGS
## 24      Minnesota  MN 52.65 44.96 51.09 47.61 43.48 31.85 46.50 42.56  DEM
## 25    Mississippi  MS 43.79 55.29 39.76 59.45 40.77 49.68 48.09 49.42  REP
## 26       Missouri  MO 44.38 53.76 46.10 53.30 44.07 33.92 44.35 51.16  REP
## 27        Montana  MT 41.70 55.35 38.56 59.07 37.63 35.12 32.43 56.82  REP
## 28       Nebraska  NE 38.03 59.80 32.68 65.90 29.40 46.58 26.04 65.53  REP
## 29         Nevada  NV 52.36 45.68 47.88 50.47 37.36 34.73 26.89 62.54  BGS
## 30  New Hampshire  NH 51.98 46.40 50.24 48.87 38.91 37.69 28.35 57.74  BGS
## 31     New Jersey  NJ 58.38 40.59 52.92 46.24 42.95 40.58 38.56 51.97  DEM
## 32     New Mexico  NM 52.99 42.84 49.05 49.84 45.90 37.34 36.78 54.97  BGS
## 33       New York  NY 63.35 35.17 58.37 40.08 49.73 33.88 43.99 46.66  DEM
## 34 North Carolina  NC 48.35 50.39 43.58 56.02 42.65 43.44 47.18 49.30  REP
## 35   North Dakota  ND 38.69 58.32 35.50 62.86 32.18 44.22 26.26 64.23  REP
## 36           Ohio  OH 50.67 47.69 48.71 50.81 40.18 38.35 40.91 51.51  BGS
## 37       Oklahoma  OK 33.23 66.77 34.43 65.57 34.02 42.65 34.97 60.50  REP
## 38         Oregon  OR 54.24 42.15 51.35 47.19 42.48 32.53 38.67 48.33  DEM
## 39   Pennsylvania  PA 51.97 46.59 50.92 48.42 45.15 36.13 42.48 49.59  BGS
## 40   Rhode Island  RI 62.70 35.24 59.42 38.67 47.04 29.02 47.67 37.20  DEM
## 41 South Carolina  SC 44.09 54.56 40.90 57.98 39.88 48.02 48.04 49.57  REP
## 42   South Dakota  SD 39.87 57.89 38.44 59.91 37.14 40.66 31.69 60.53  REP
## 43      Tennessee  TN 39.08 59.48 42.53 56.80 47.08 42.43 48.41 48.70  REP
## 44          Texas  TX 41.38 57.17 38.22 61.09 37.08 40.56 41.42 55.28  REP
## 45           Utah  UT 24.75 72.79 26.00 71.54 24.65 43.36 20.57 72.77  REP
## 46        Vermont  VT 66.57 30.97 58.94 38.80 46.11 30.42 38.41 44.37  DEM
## 47       Virginia  VA 51.16 47.28 45.48 53.68 40.59 44.97 40.31 53.03  BGS
## 48     Washington  WA 56.16 41.29 52.82 45.64 43.41 31.97 37.32 49.66  DEM
## 49 West Viriginia  WV 35.54 62.30 43.20 56.06 48.41 35.39 49.81 45.30  BGS
## 50      Wisconsin  WI 52.83 45.89 49.70 49.32 41.13 36.78 43.18 47.90  BGS
## 51        Wyoming  WY 27.82 68.64 29.07 68.86 34.10 39.70 27.97 62.64  REP
##         REG
## 1     South
## 2      West
## 3      West
## 4     South
## 5      West
## 6      West
## 7  Northest
## 8     South
## 9     South
## 10    South
## 11    South
## 12     West
## 13     West
## 14  Midwest
## 15  Midwest
## 16  Midwest
## 17  Midwest
## 18    South
```

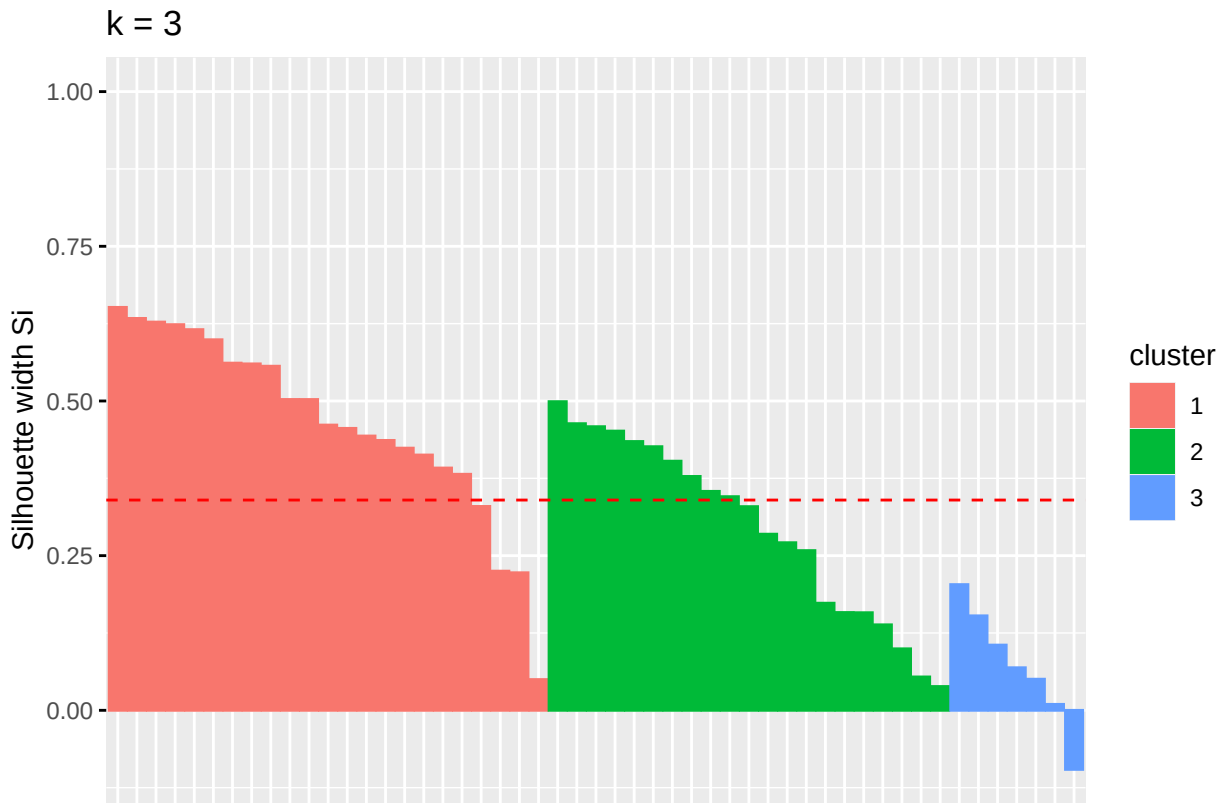
```
## 19    South
## 20 Northest
## 21    South
## 22 Northest
## 23  Midwest
## 24  Midwest
## 25    South
## 26  Midwest
## 27     West
## 28  Midwest
## 29     West
## 30 Northest
## 31 Northest
## 32     West
## 33 Northest
## 34    South
## 35  Midwest
## 36  Midwest
## 37    South
## 38     West
## 39 Northest
## 40 Northest
## 41    South
## 42  Midwest
## 43    South
## 44    South
## 45     West
## 46 Northest
## 47    South
## 48     West
## 49    South
## 50  Midwest
## 51     West
```



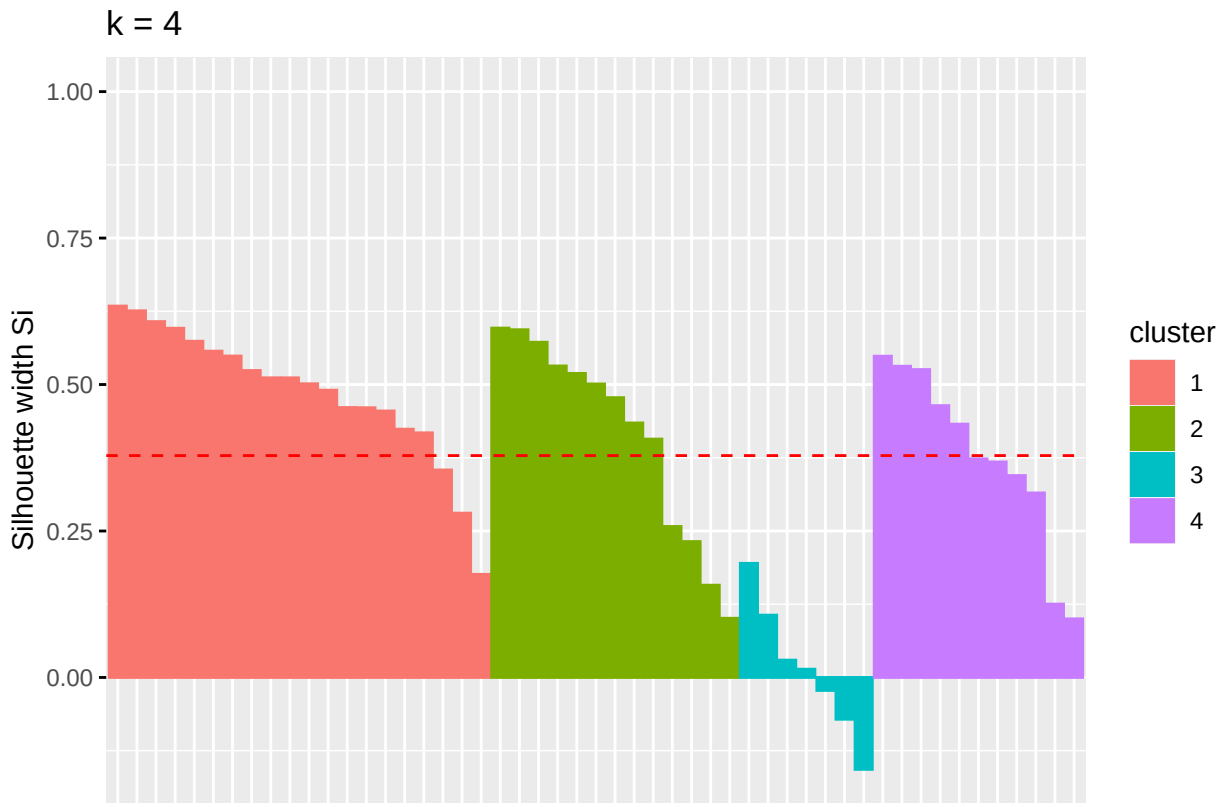

```
##   cluster size ave.sil.width
## 1       1   21          0.46
## 2       2   30          0.36
```



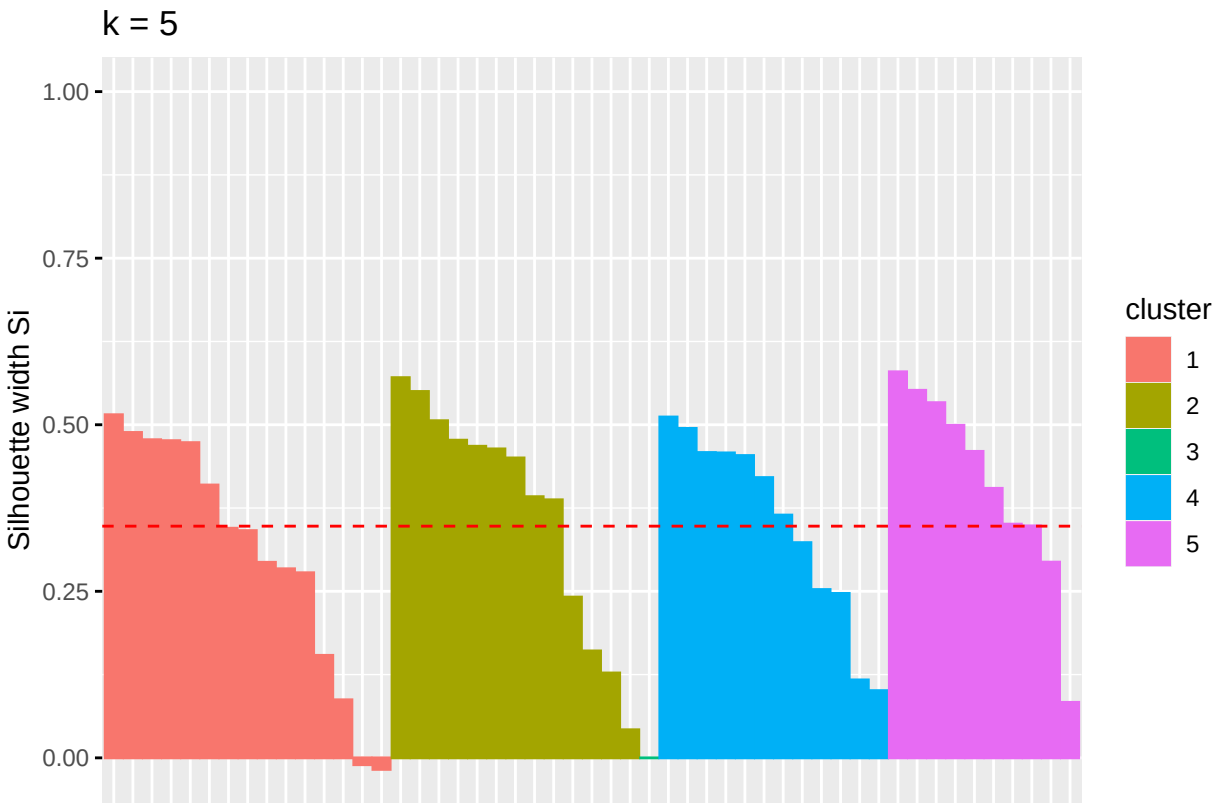

```
##   cluster size ave.sil.width
## 1       1   23          0.46
## 2       2   21          0.29
## 3       3    7          0.07
```

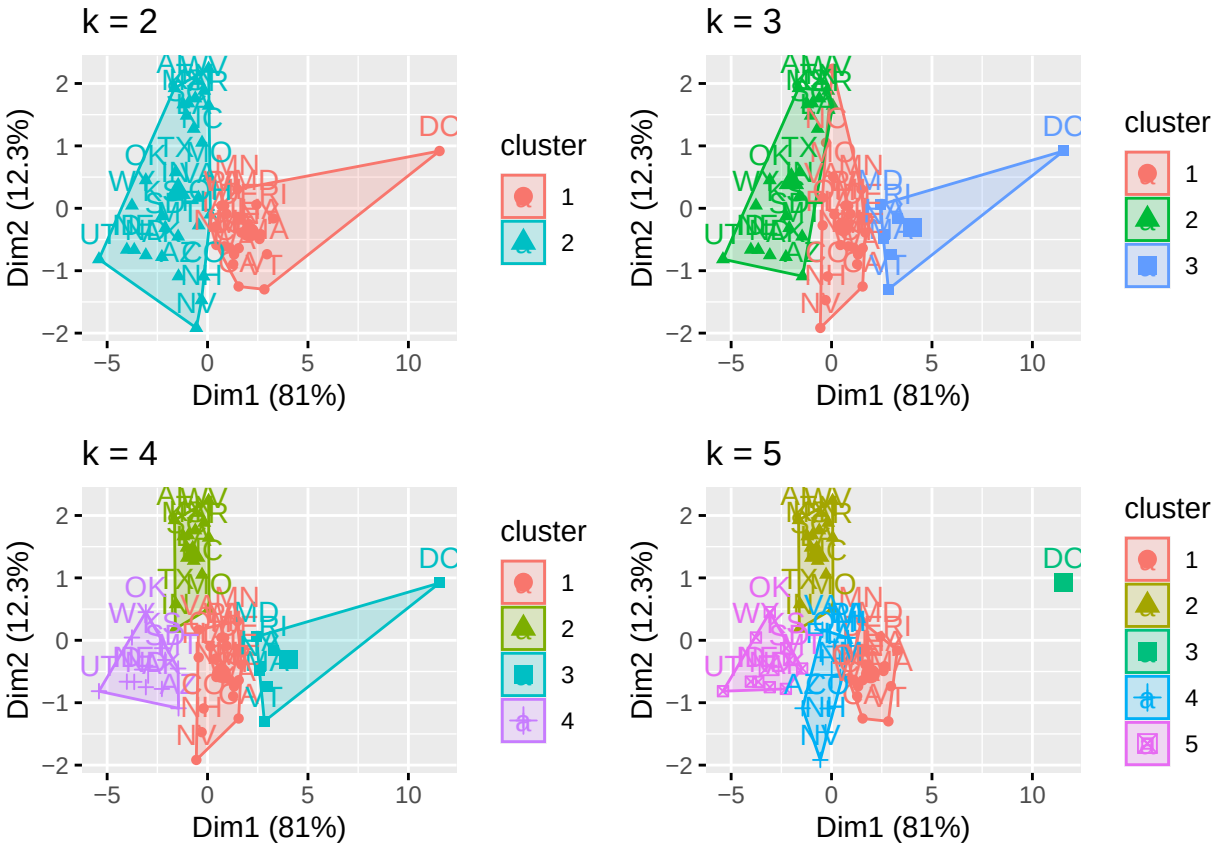


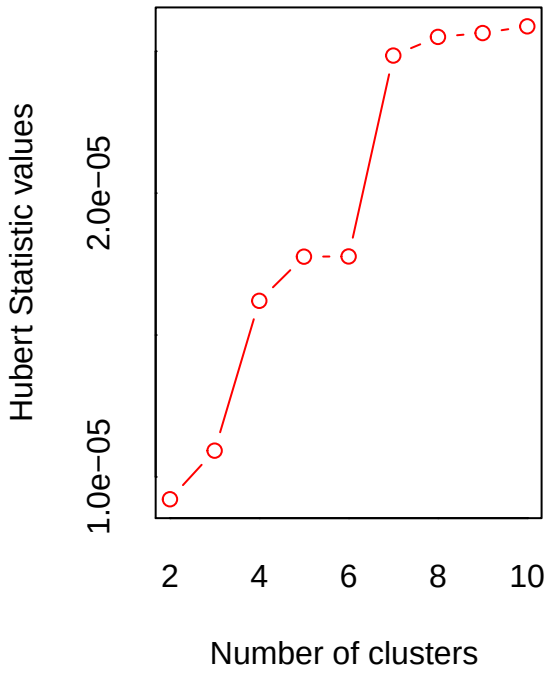

```
##   cluster size ave.sil.width
## 1       1   20          0.49
## 2       2   13          0.41
## 3       3    7          0.01
## 4       4   11          0.38
```




```
##   cluster size ave.sil.width
## 1       1   15          0.31
## 2       2   13          0.37
## 3       3    1          0.00
## 4       4   12          0.35
## 5       5   10          0.41
```

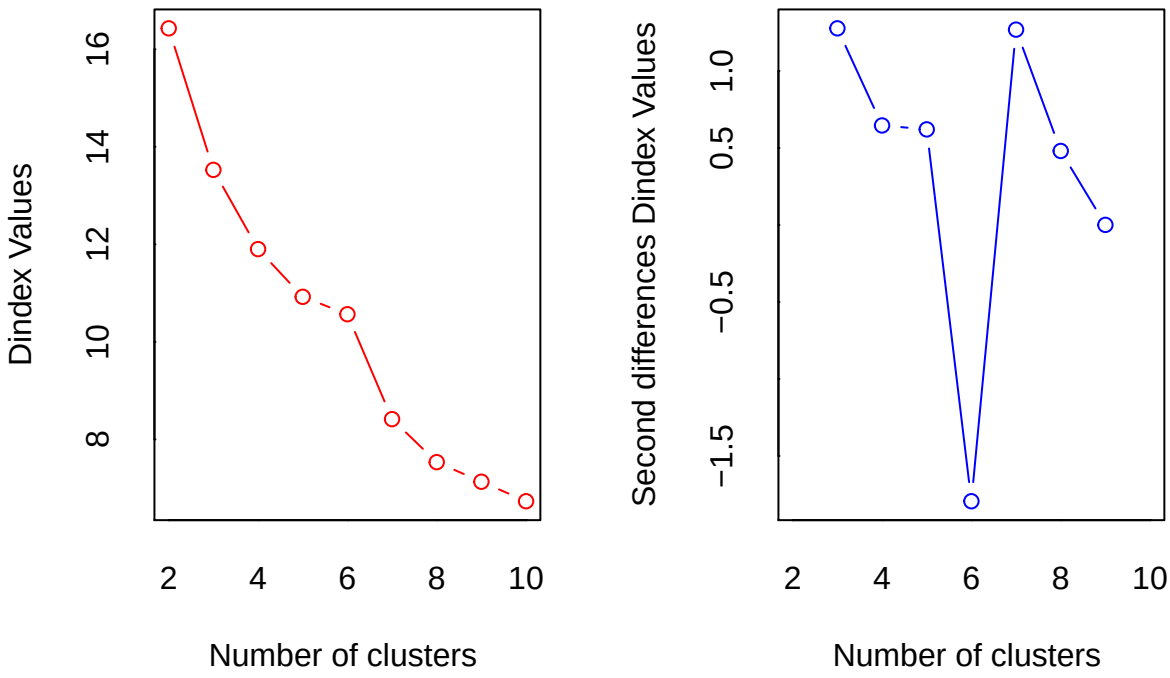

k = 5
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
4
5


k = 2
k = 3
k = 4
k = 5
Dim1 (81%)
Dim2 (12.3%)
cluster
DC



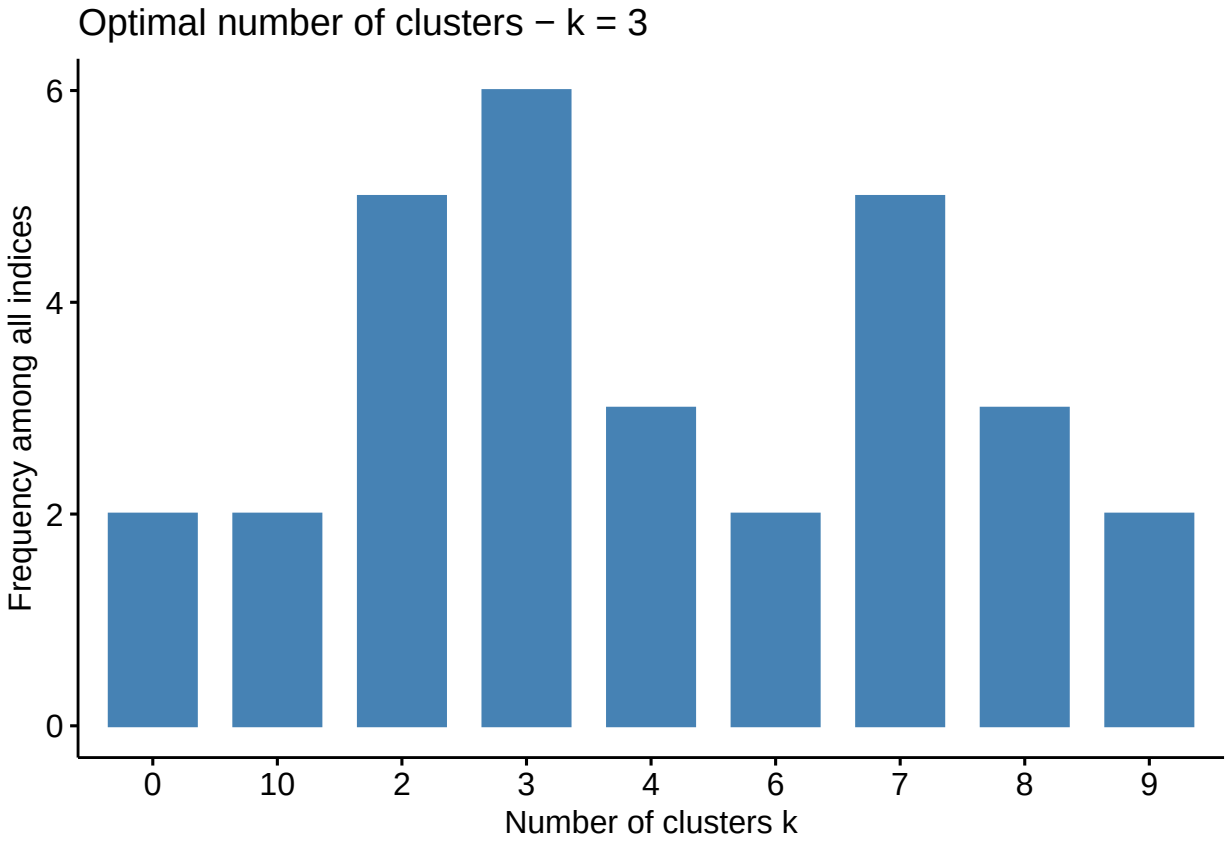
Ahora nos dedicaremos a averiguar ¿Cuántos Clusters deberíamos usar?

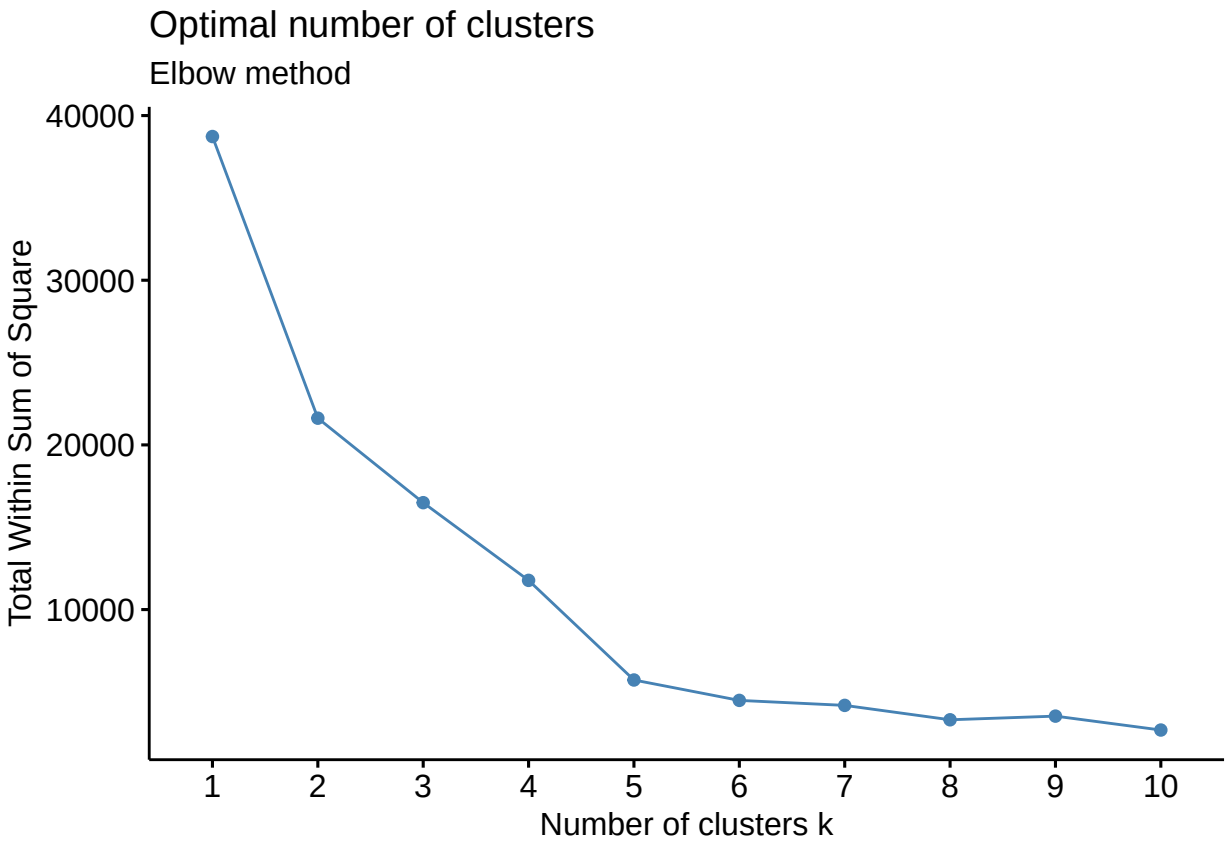
```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```

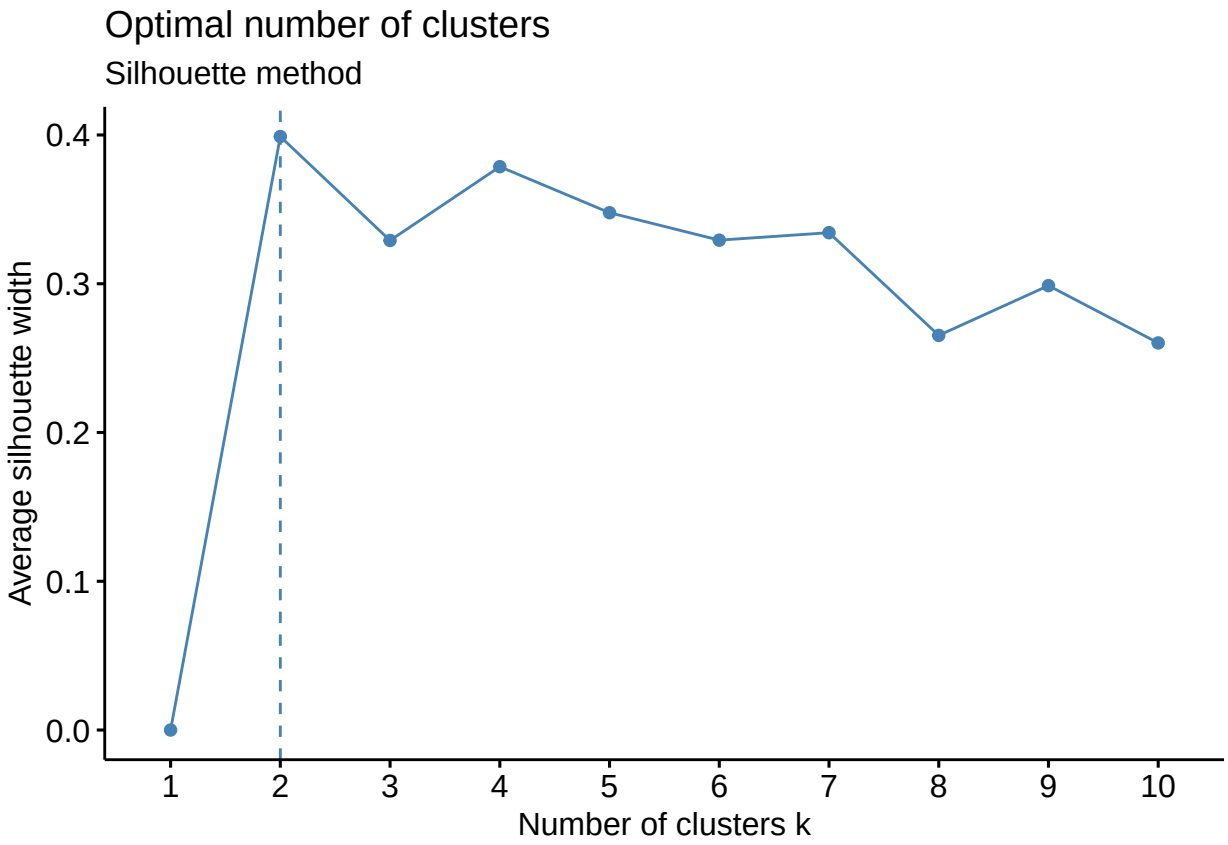


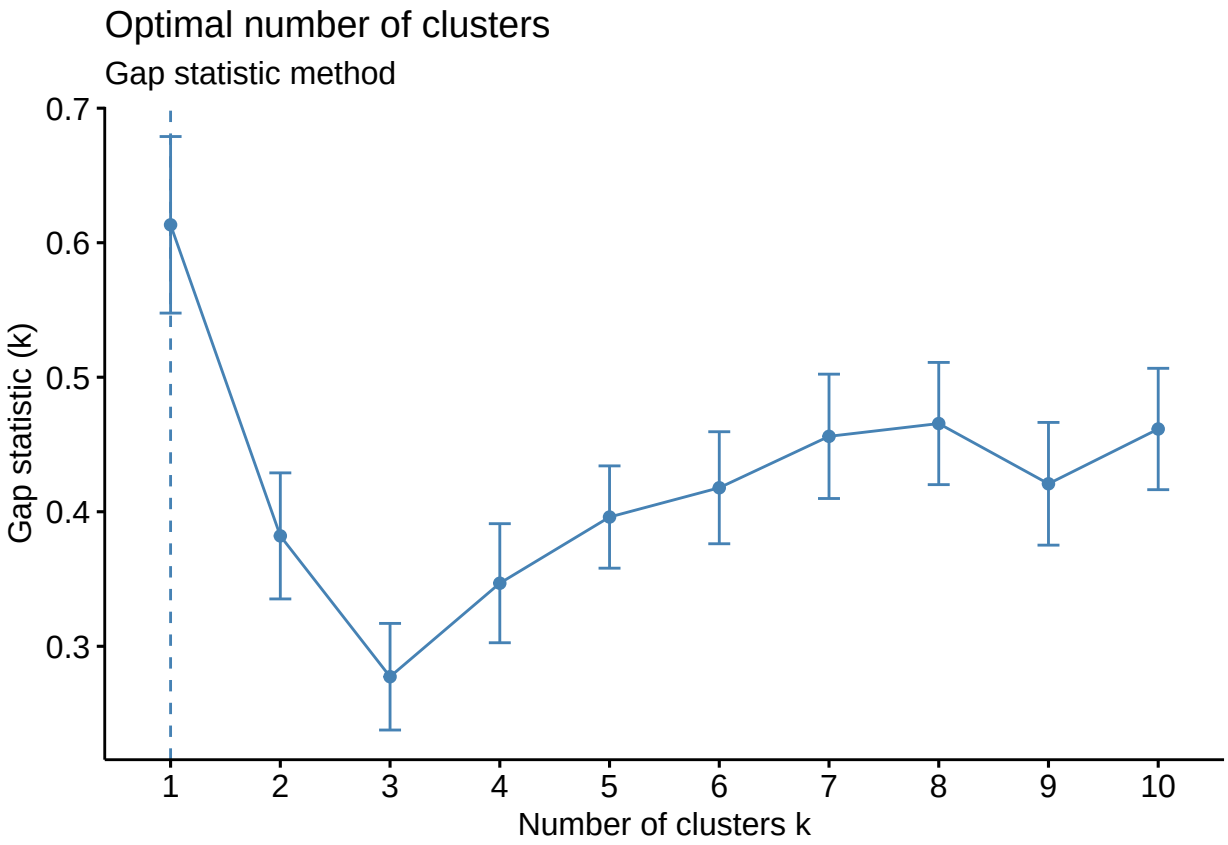

```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 5 proposed 2 as the best number of clusters
## * 6 proposed 3 as the best number of clusters
## * 3 proposed 4 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 5 proposed 7 as the best number of clusters
## * 3 proposed 8 as the best number of clusters
## * 2 proposed 9 as the best number of clusters
## * 2 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************

## Among all indices:
## ==================
## * 2 proposed  0 as the best number of clusters
## * 5 proposed  2 as the best number of clusters
```

```
## * 6 proposed  3 as the best number of clusters
## * 3 proposed  4 as the best number of clusters
## * 2 proposed  6 as the best number of clusters
## * 5 proposed  7 as the best number of clusters
## * 3 proposed  8 as the best number of clusters
## * 2 proposed  9 as the best number of clusters
## * 2 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  3 .
```

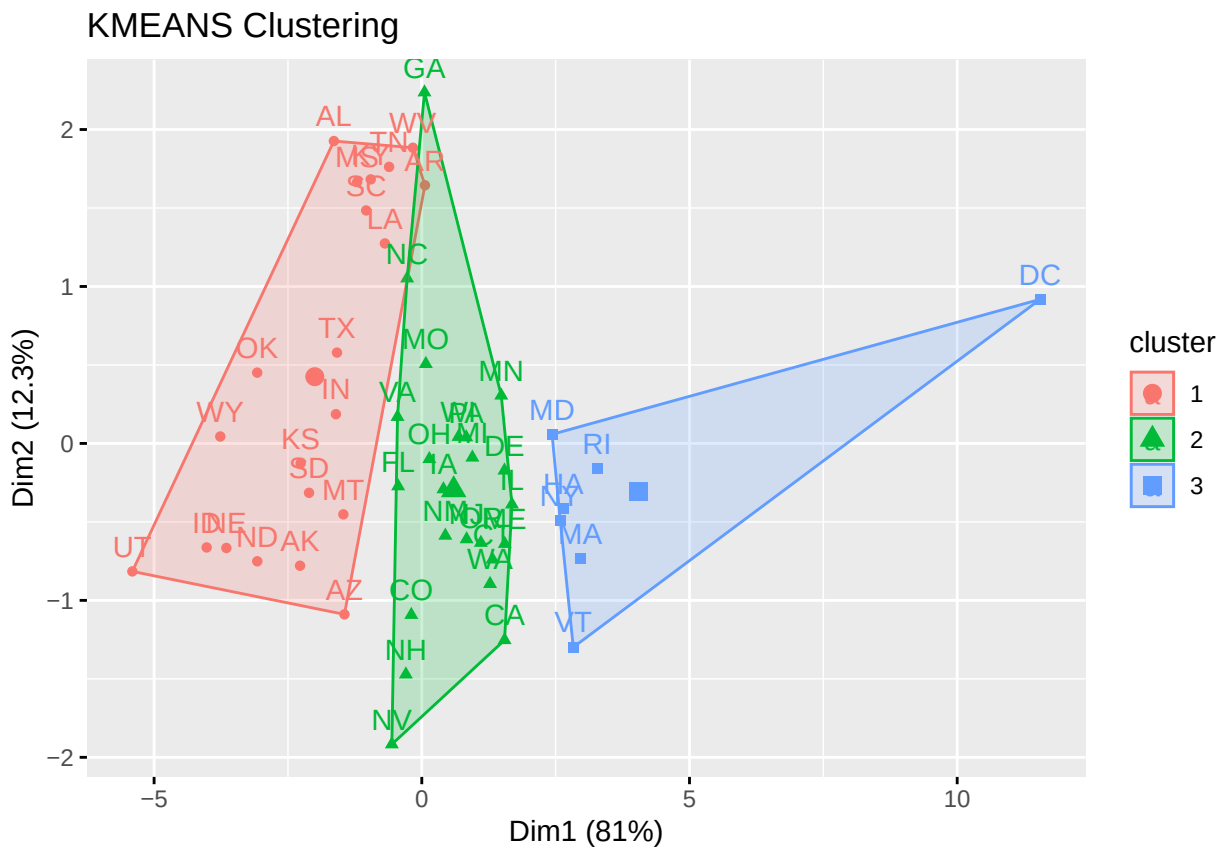



Optimal number of clusters
Elbow method
Total Within Sum of Square
40000
30000
20000
10000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

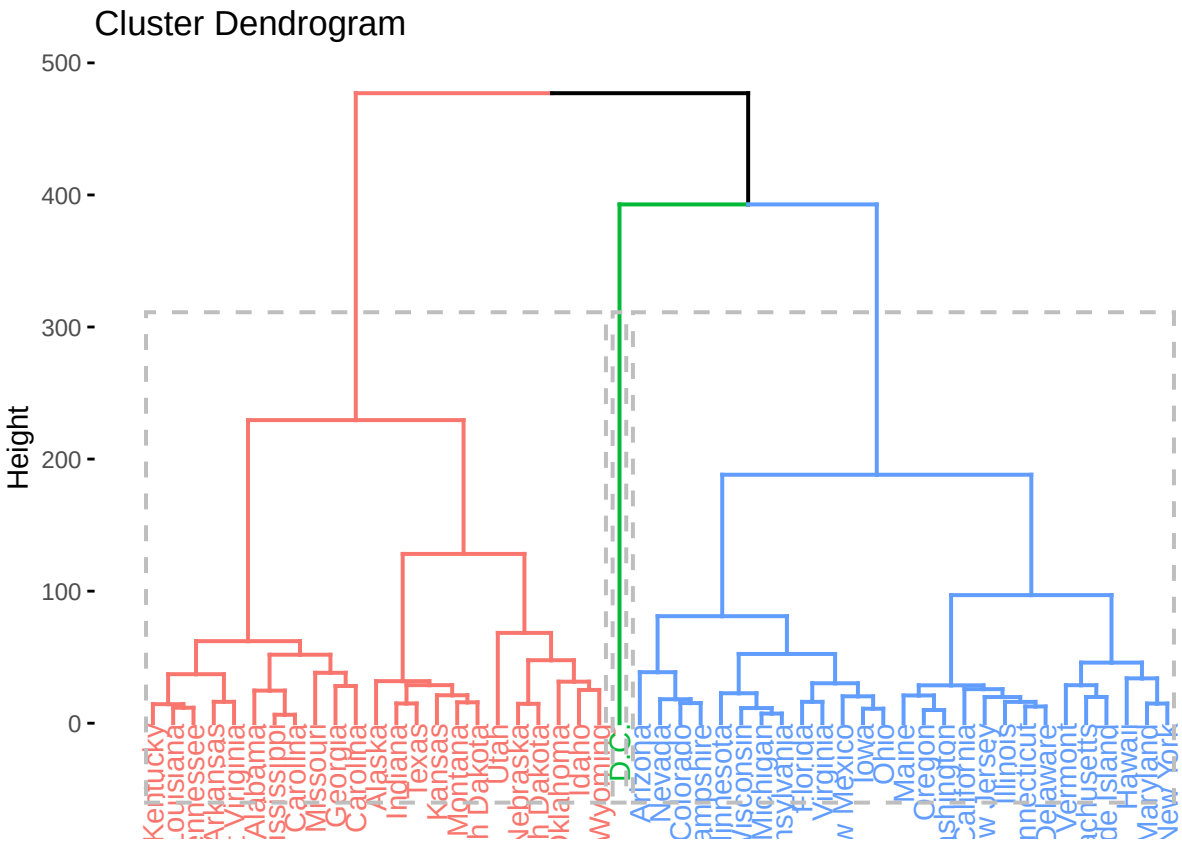

Optimal number of clusters
Silhouette method
Average silhouette width
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Como podemos ver, algunos criterio sugieren que un cluster es suficiente, lo cual anularía el propósito de clusterización, sin embargo la mayoría de los criterios sugieren que K = 3 es el número óptimo de clusterización.



Realizando ahora, por el método jerárquico



En el dendograma podemos ver como podrían dividirse los estados. A diferencia del método de K-Medias, elmétodo jerárquico úbica al DC en un grupo separado de los demás estados, esto posiblemente debido a la prominente diferencia e inclinación hacia el partido demócrata de éste distrito, cosa que se puede apreciar mejor en el HeatMap anterior.

Por lo que el modelo propuesto por el método jerárquico podría ser el mejor para usar.

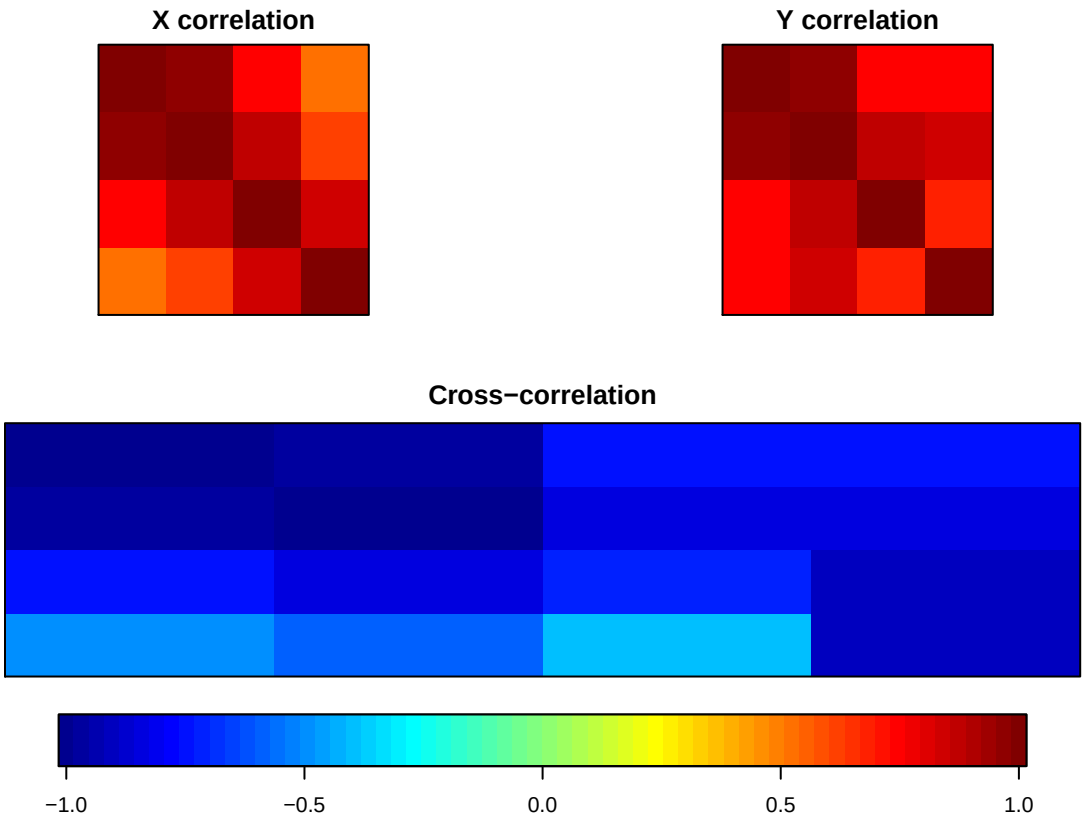
3. Análisis de Correlación Canónica: Realice un análisis de correlación canónica de las variables que se refieren a los porcentajes de votación por el partido repúblicano contra las variables que se refieren a los porcentajes de votación por el partido demócrata. Escriba cuales son los modelos lineales que obtendría, y cuales son las correlaciones entre las nuevas variables y las variables originales.

Comenzamos separándo los datos por partido

```
demo <- data %>%
  select(DEM12,DEM04,DEM92,DEM80)
gop <- data %>%
  select(GOP12,GOP04,GOP92,GOP80)
```

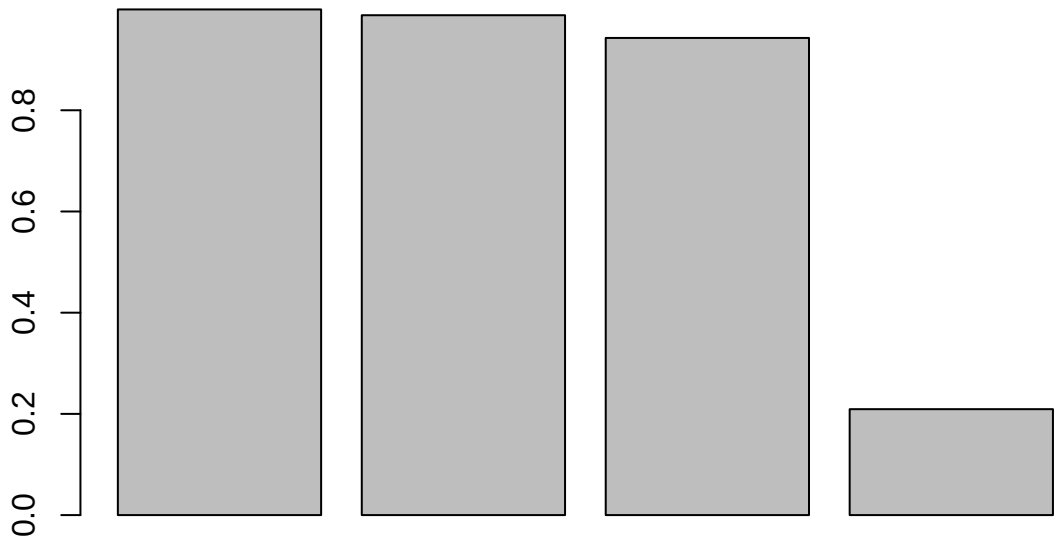
Luego cálculamos la correlación entre las nuevas variables y la correlación cruzada.

```
correl <- matcor(demo,gop)
img.matcor(correl, type = 2)
```



Como era de esperarse, la correlación cruzada es principalmente negativa, el que crezca el porcentaje del GOP disminuye el de los demócratas.

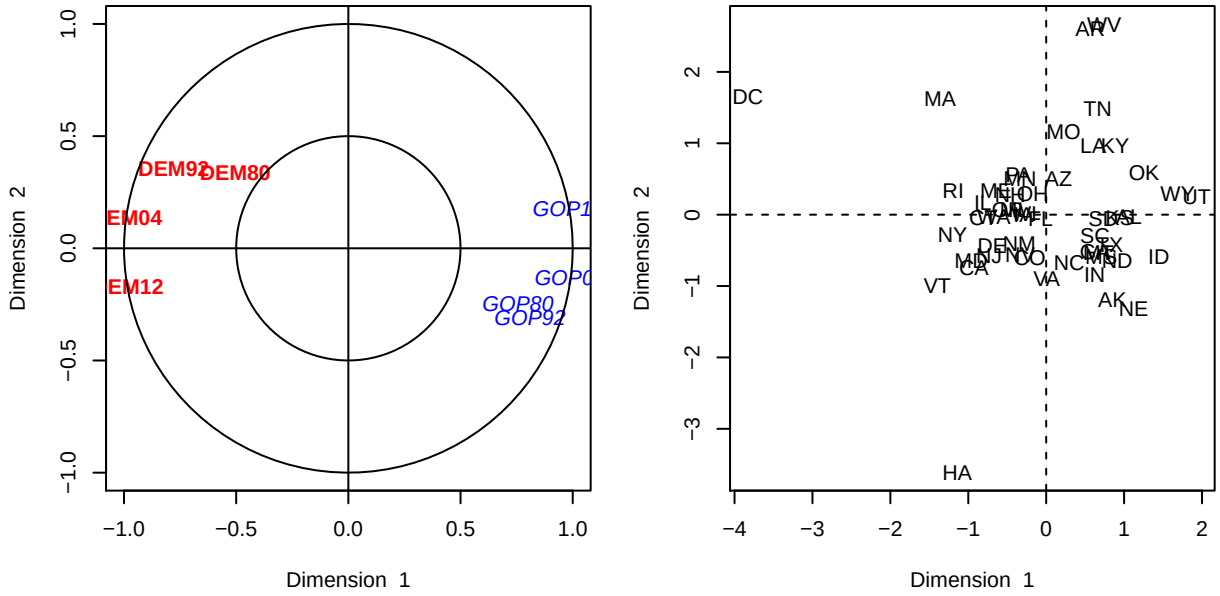
```
cc1 <- cc(demo,gop)
barplot(cc1$cor, main = "Canonical correlations for 'cancor()'", col = "gray")
```




```
cc1$xcoef
```

```
##             [,1]     [,2]     [,3]    [,4]
## DEM12 -0.032436 -0.27647  0.08360 -0.1155
## DEM04 -0.063832  0.31917 -0.14135  0.3035
## DEM92 -0.004069 -0.01908 -0.03723 -0.3683
## DEM80  0.012221  0.01259  0.15419  0.1413
```

```
plt.cc(cc1, var.label = TRUE, ind.names = data[,2])
```



Nuevamente podemos ver que el DC parece ser atípico entre los demás estados y como las variables que representan los resultados de los demócratas y repúblicanos , parecen estar correlacionadas entre sí mismas mas que entre ambos partidos. Sin embargo vemos como ningún año en específico es más representativo que los demás para la variabilidad de los datos.

4. Análisis de Correspondencias:

Comenzamos por añadir una forma de dividir los datos según lo pedido por el enunciado.

- A = > 10% a favor de repúblicanos.
- B Entre 2% y 10% a favor de repúblicanos.
- C < 2% de diferencia entre los dos partidos.
- D Entre 2% y 10% a favor de demócratas.
- E > 10% a favor de demócratas.

Comenzamos añadiendo las nuevas variables a nuestro data set por año, resultando:

```
##       Estado Cod DEM12 GOP12 DEM04 GOP04 DEM92 GOP92 DEM80 GOP80 Tend   REG Y80
## 1    Alabama  AL 38.36 60.55 36.84 62.46 40.88 47.65 47.45 48.75  REP South   C
## 2     Alaska  AK 40.81 54.80 35.52 61.07 30.29 39.46 26.41 54.35  REP  West   E
## 3    Arizona  AZ 44.59 53.65 44.40 54.87 36.52 38.47 28.24 60.61  REP  West   E
## 4   Arkansas  AR 36.88 60.57 44.55 54.31 53.21 35.48 47.52 48.13  REP South   C
## 5 California  CA 60.24 37.12 54.31 44.36 46.01 32.61 35.91 52.69  DEM  West   E
## 6   Colorado  CO 51.49 46.13 47.02 51.69 40.13 35.87 31.07 55.07  BGS  West   E
##   Y92 Y04 Y12
## 1   D   E   E
```

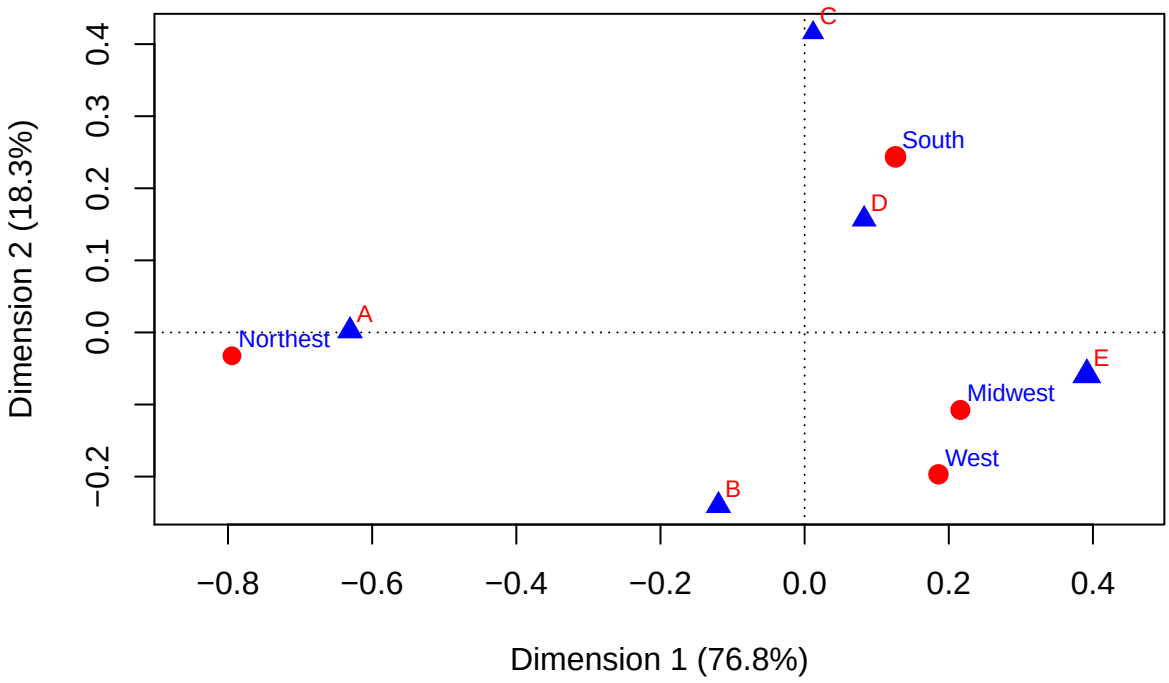
```
## 2   D   E   E
## 3   C   E   D
## 4   A   D   E
## 5   A   B   A
## 6   B   D   B
```

Creamos la tabla de contingencia contando las ocurrencias de A,B,c,D y E por región

```
##           Y80
## REG         A  B  C  D  E
##   Midwest   5 11  3 10 19
##   Northest 17  9  3  5  2
##   South    11  8 10 14 25
##   West      8 11  3  6 24
```

Así creamos realizamos nuestro análisis de correspondencias

```
library('ca')
rel.ca <- ca(tablea)
plot.ca(rel.ca,mass=c(T,T),col = c('red','blue'))
```

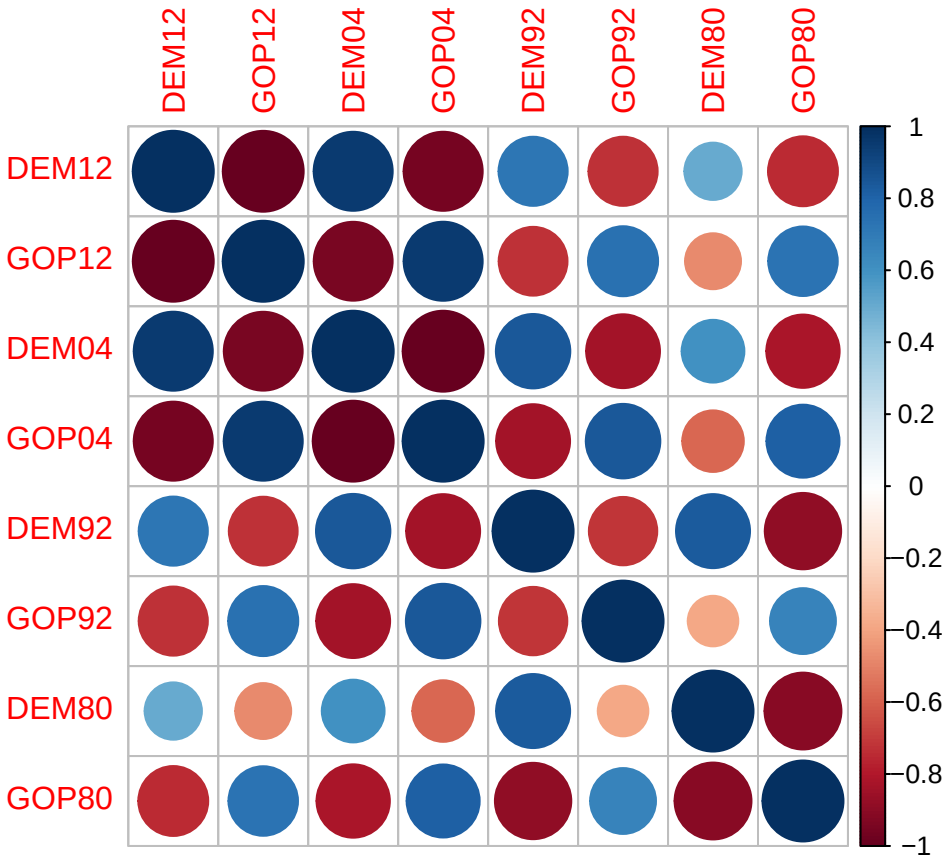


Como podemos ver los estados de la región *South* tienen una tendencia a favorecer, por un bajo margen, a los demócratas. Por otro lado las zonas de *Midwest* y *WEST* suelen ser de mayoría democrática. La región *Northest* tiende a ser más favorecedora del GOP.

6. Análisis Factorial:Realice un análisis factorial sobre los datos. Consiga el número de factores que serían suficientes para explicar los datos. Escriba los modelos de cada uno de los factores, y los puntajes factoriales que se obtendrían.

Comenzamos calculando la matriz de covarianzas de los datos

```
cor.usa <- cor(usa)
corrplot(cor.usa)
```



Realizamos algunas pruebas a los datos

```
KMO(usa)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = usa)
## Overall MSA =  0.77
## MSA for each item =
## DEM12 GOP12 DEM04 GOP04 DEM92 GOP92 DEM80 GOP80
##  0.74  0.74  0.77  0.77  0.85  0.92  0.63  0.76
```

```
det(cor.usa)
```

```
## [1] 1.5e-10
```

```
bartlett.test(usa)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  usa
```

```
## Bartlett's K-squared = 25, df = 7, p-value = 9e-04
```

Vemos que el KMO es de un valor mediano. Por otro lado el determinante de la matriz de correlación es prácticamente 0, lo que indíca multicolinealidad. Adicionalmente el teste de Barlett rechaza la hipótesis de que los datos vienen de una distribución por lo cuál el análisis factorial puede no ser aceptable

Como factor para elegir el número de factores considramos el número de componentes principales escogidas en los ejercicios anteriores

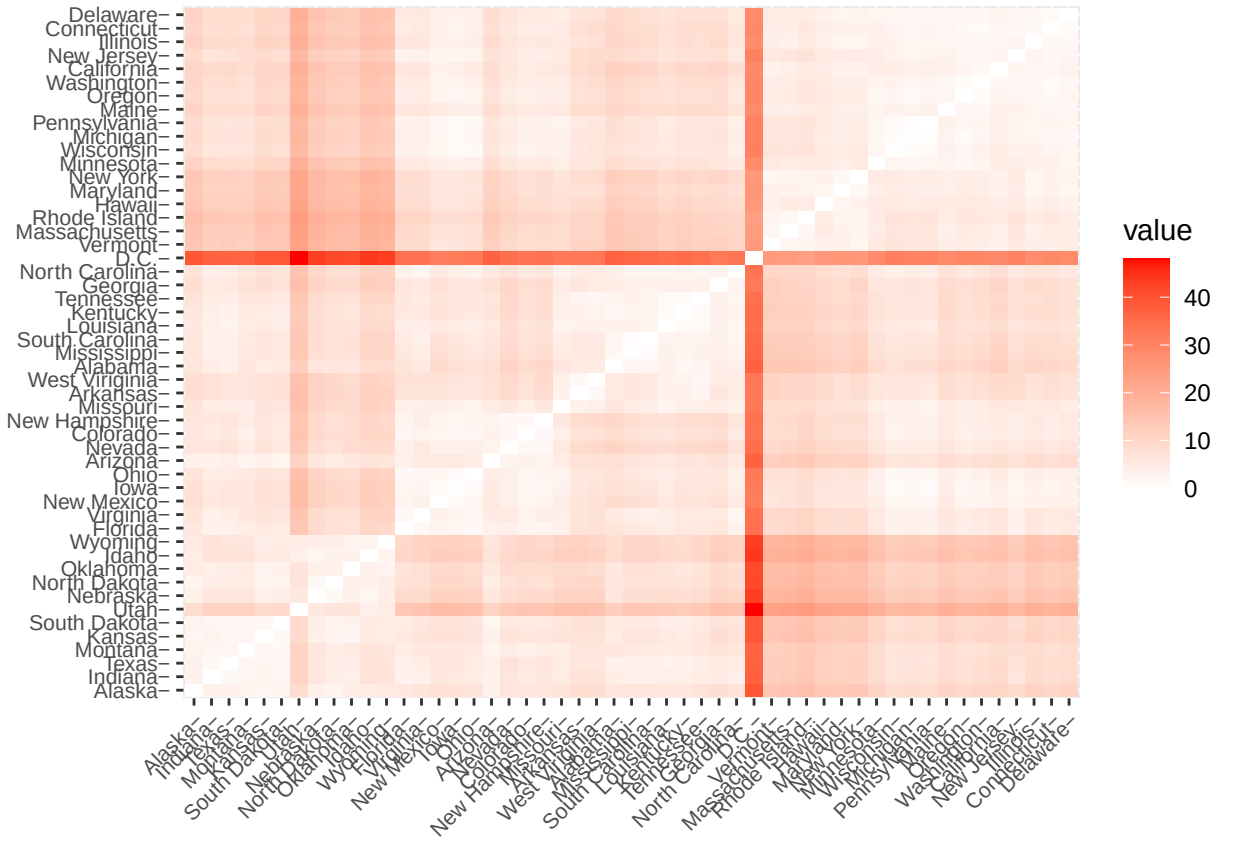
```
factanal(usa, factors = 3, rotations = "none")
```

```
##
## Call:
## factanal(x = usa, factors = 3, rotations = "none")
##
## Uniquenesses:
## DEM12 GOP12 DEM04 GOP04 DEM92 GOP92 DEM80 GOP80
## 0.005 0.005 0.005 0.005 0.098 0.201 0.005 0.059
##
## Loadings:
##       Factor1 Factor2 Factor3
## DEM12 -0.905   0.314  -0.282
## GOP12  0.910  -0.288   0.292
## DEM04 -0.745   0.421  -0.514
## GOP04  0.753  -0.396   0.524
## DEM92 -0.417   0.731  -0.441
## GOP92  0.518  -0.231   0.690
## DEM80 -0.181   0.975  -0.106
## GOP80  0.462  -0.808   0.273
##
##                Factor1 Factor2 Factor3
## SS loadings      3.455   2.707   1.461
## Proportion Var   0.432   0.338   0.183
## Cumulative Var   0.432   0.770   0.953
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 116 on 7 degrees of freedom.
## The p-value is 5.14e-22
```

Como podemos ver, con tres factores podemos describir el 95% de la varianza.

7. Escalamiento Multidimensional: Encuentre una distancia entre los estados a partir de los datos. Utilizando escalamiento multidimensional, realizar un mapa de los estados + el Distrito de Columbia. ¿Qué agrupaciones podríia realizar?

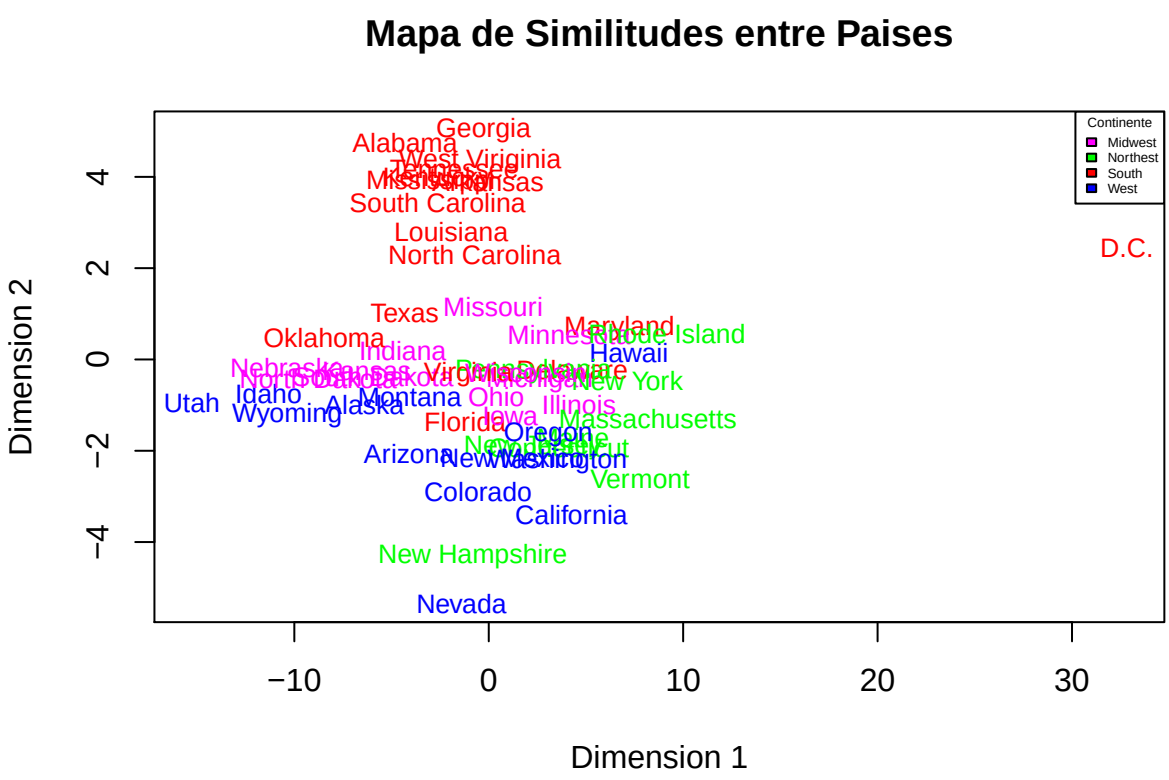
Comenzamos calculando la matriz de distancias entre variables, en este caso utilizaremos la distancia de *Manhattan*



Cómo podemo ver el Distrito de Columbia es el más alejado de los demás.

Ahora realizamos el escalamiénto multidimensional para dos dimensiones

```
## [1] 0.8266 0.9030
```



Como podemos ver, la agrupación conseguida puede verse similar a la determinada por regiones, a excepción del DC los estados de las mismas regiones tienden a estar más cerca entre sí, en especial los estados del *South*