Actividad 2Estadística Inferencial Ingeniería Industrial Profesora: Iliana Paternina Ortega

Observaciones:

- **Todo lo realizado debe ser justificado de manera clara y precisa, con argumentos estadísticos, sino no se le dará valoración.**
- **La actividad puede ser entregada en parejas**
- **Cualquier intento de plagio y/o fraude será sancionado con una clasificación de 0.0**

Nombre Completo: Braikan piña salcedo, Steven gomez foliaco Fecha: 10 / 09 / 2023

Análisis de Clúster (Grupo2)

1. Para realizar la Actividad 2, debes crear un Rscript en RStudio y guardar el archivo con un nombre específico de tu preferencia. Este debe ser colgado en SAVI. Coloca nombres en el encabezado del script.

2. Para realizar la Actividad 2, debes crear un Word y guardar el archivo con un nombre específico de tu preferencia. Este debe ser colgado en SAVI. Coloca nombres en el encabezado del doc.

3. Carga los datos o dataset llamado **"States"** que se encuentran en el paquete **"carData"**

El conjunto de datos de States corresponde a un estudio en 51 estados estadounidenses.
Las variables medidas fueron:
región
pop: en miles.
SATV
Puntaje promedio de los estudiantes que se gradúan de secundaria en el estado en el componente verbal de la Prueba de Aptitud Escolar (un examen estándar de admisión a la universidad).
SATM
Puntaje promedio de los estudiantes graduados de secundaria en el estado en el componente de matemáticas de la Prueba de Aptitud Escolar.
Percent: Porcentaje de estudiantes graduados de secundaria en el estado que tomaron el examen SAT.
Dollars:Gasto estatal en educación pública, en miles de dólares por estudiante.
Pay: Salario promedio de un maestro en el estado, en miles de dólares.

Se quiere encontrar grupos de comportamiento similares según un análisis clúster, a través de los métodos jerárquicos aglomerativos, utilizando la distancia de Manhattan con el método del vecino más lejano.

Para realizar el anterior análisis debes seguir los pasos:

```
install.packages("carData")
library(carData)
```

```
# Conociendo el dataset

help("States")
str(States)
summary(States)

# Punto 3
data(States)
```

4. Crea un conjunto de datos nuevo con sólo variables numéricas y omite los valores NA´s que pueda haber.

```
# Punto 4
datos <- States[-1]

# Se omiten los NA's
datos <- na.omit(datos)
```

5. Revisa la escala de cada variable dentro de tu nuevo dataset y decide si debes aplicar la función scale. **Explica**.

```
> summary(datos)
      pop             SATV            SATM          percent         dollars
 Min.   :  454   Min.   :397.0   Min.   :437.0   Min.   : 4.00   Min.   :2.993
 1st Qu.: 1215   1st Qu.:422.5   1st Qu.:470.0   1st Qu.:11.50   1st Qu.:4.354
 Median : 3294   Median :443.0   Median :490.0   Median :25.00   Median :5.045
 Mean   : 4877   Mean   :448.2   Mean   :497.4   Mean   :33.75   Mean   :5.175
 3rd Qu.: 5780   3rd Qu.:474.5   3rd Qu.:522.5   3rd Qu.:57.50   3rd Qu.:5.689
 Max.   :29760   Max.   :511.0   Max.   :577.0   Max.   :74.00   Max.   :9.159
      pay
 Min.   :22.00
 1st Qu.:27.50
 Median :30.00
 Mean   :30.94
 3rd Qu.:33.50
 Max.   :43.00
```

Es necesario escalar los datos debido a que tienen escalas diferentes, y resulta difícil comparar las variables directamente, es por esa razón que escalamos los datos para tener una escala “común” por así decirlo entre ellas facilitando el análisis a lo largo de este trabajo

Aplicamos la función scale

```
summary(datos)
datos2 <- scale(datos)
```

6. Calcula las distancias de manhattan.

```
d <- dist(datos2, method = "manhattan")
```

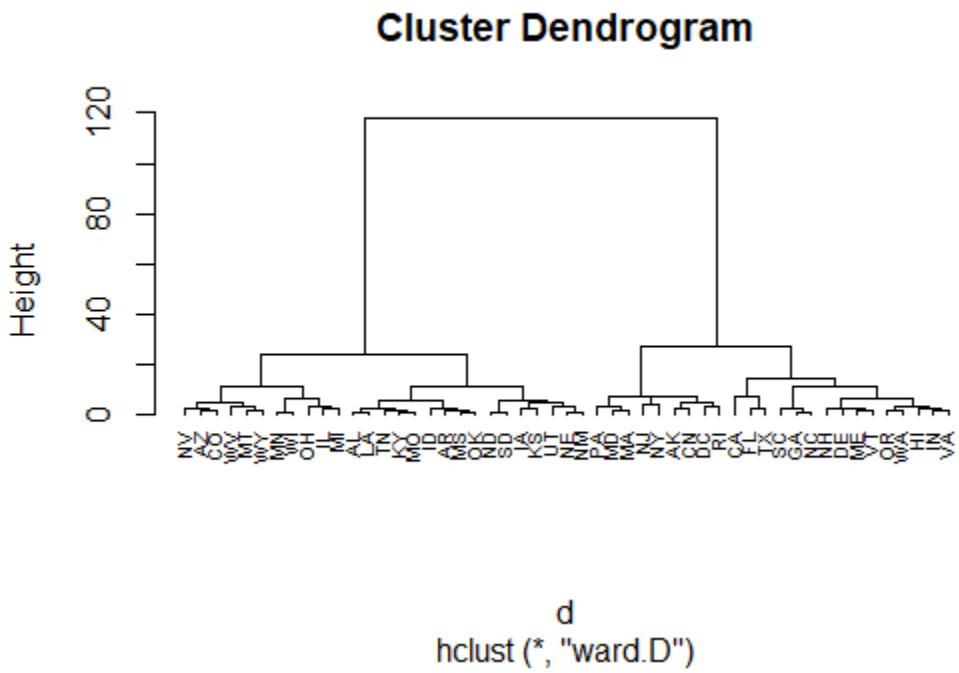
7. Halla el coeficiente de aglomeración. Interpreta

Se aplicó el método de ward debido a que se quería agrupar datos en clústeres de tal manera que la varianza dentro de cada clúster sea lo más pequeña posible.

```
> hc2 <- agnes(d, method = "ward")
> coef_aglomeracion <- hc2$ac
> coef_aglomeracion
[1] 0.9558226
```

El coeficiente de aglomeración es igual a 0.9558, lo que indica una fuerte aglomeración en los datos ya que se aproxima a 1. Esto significa que estos datos tienden a formar grupos bien definidos en lugar de estar dispersos.

8. Debes mostrar el dendograma, explicar lo que se vislumbra en torno a los clústeres. Corta el dendrograma según el historial de aglomeración y comenta a qué altura lo cortas y por qué tu elección.



En torno a los clusters podemos explicar que a partir del gráfico podemos darnos una idea sobre cómo se agrupan los datos en clústeres en función de sus similitudes o distancia. A medida que bajamos en el dendrograma, los elementos se agrupan en más grupos y con el simple hecho de ver el gráfico podemos identificar el número de clústeres para tener una noción de los grupos a tomar y a su vez determinar dónde cortar.

En esta imagen no podemos apreciar todo el historial completo, sin embargo se logra focalizar lo más importante que es donde empieza a aparecer un cambio inesperado después que llegamos a una altura igual a 14.

	height	merge.1	merge.2
44	10.5086086	36	40
45	10.5409353	37	38
46	11.0873604	17	39
47	14.0434539	42	46
48	23.6085180	44	45
49	26.6775879	43	47
50	117.9175052	48	49

Según el historial de aglomeración debemos cortar en una altura 14 ya que a partir de allí empiezan a haber cambios bastantes grandes en comparación con las alturas anteriores que solo variaba de 1 unidad o menos. La razón de esta elección es porque nuestras nociones del número de clústeres deseados coinciden con los posibles grupos que pueden salir a una altura de corte en 14.

9. Aplica la regla de la mayoría y explica los resultados. Explica cuántos clusters escoges y por qué.

```
*******************************************************************
> #Punto 9
> #regla de la mayoria
> res <- NbClust(datos2, distance = "manhattan", min.nc = 4, max.nc = 7,
+                method = "ward.D", index = "alllong")
*** : 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:
* 7 proposed 4 as the best number of clusters
* 14 proposed 5 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 4 proposed 7 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  5
```

De acuerdo a los resultado de la regla de la mayoría obtenemos lo siguiente

- 7 propusieron 4 como mejor número de clusters
- 14 propusieron 5 como mejor número de clusters
- 1 propuso 6 como el mejor número de clusters
- 4 propusieron 7 como mejor número de clusters

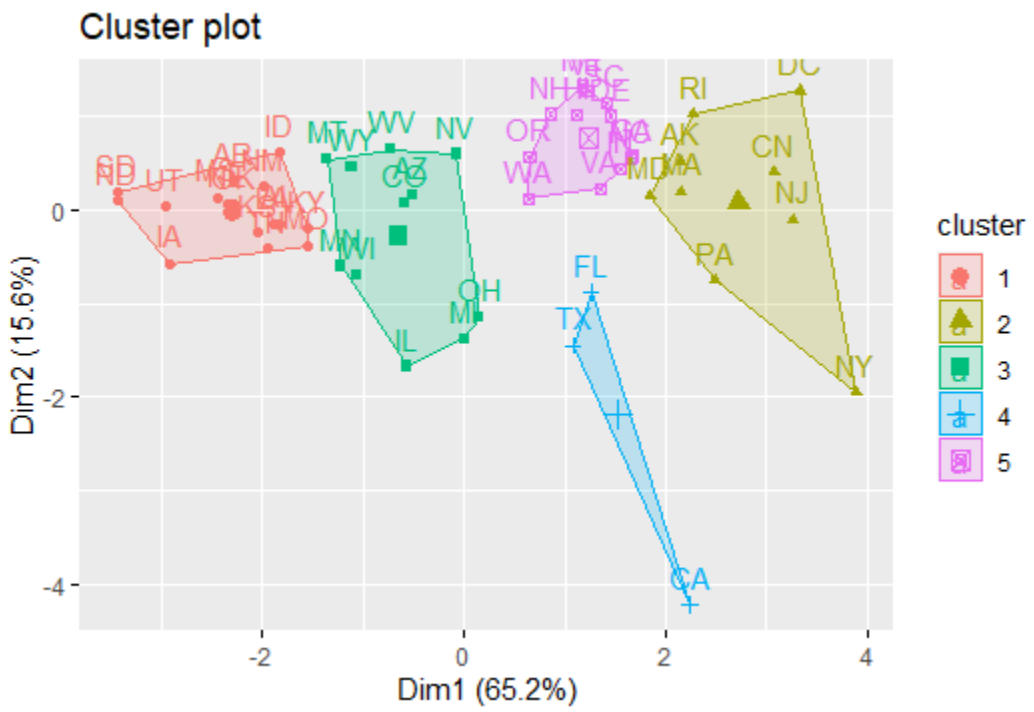
El mejor número de clusters es 5 ya que tiene más votos a favor.

Elegimos 5 clusters porque la regla de la mayoría lo sugiere y porque coincide con el número de clusters que podrían salir después de hacer el corte en la altura especificada y explicada en la pregunta anterior.

10. Muestra el gráfico de pertenencia de individuos al clúster. Fija el número de clúster según la regla de la mayoría.

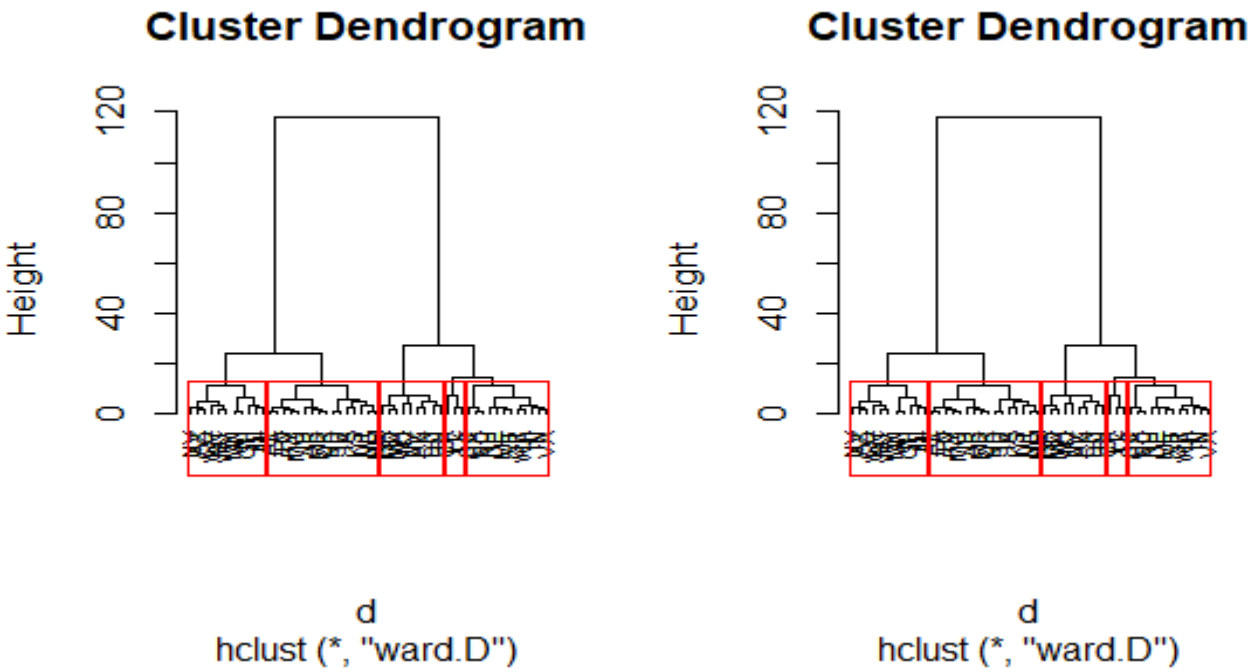
Fijamos 5 clústeres y graficamos

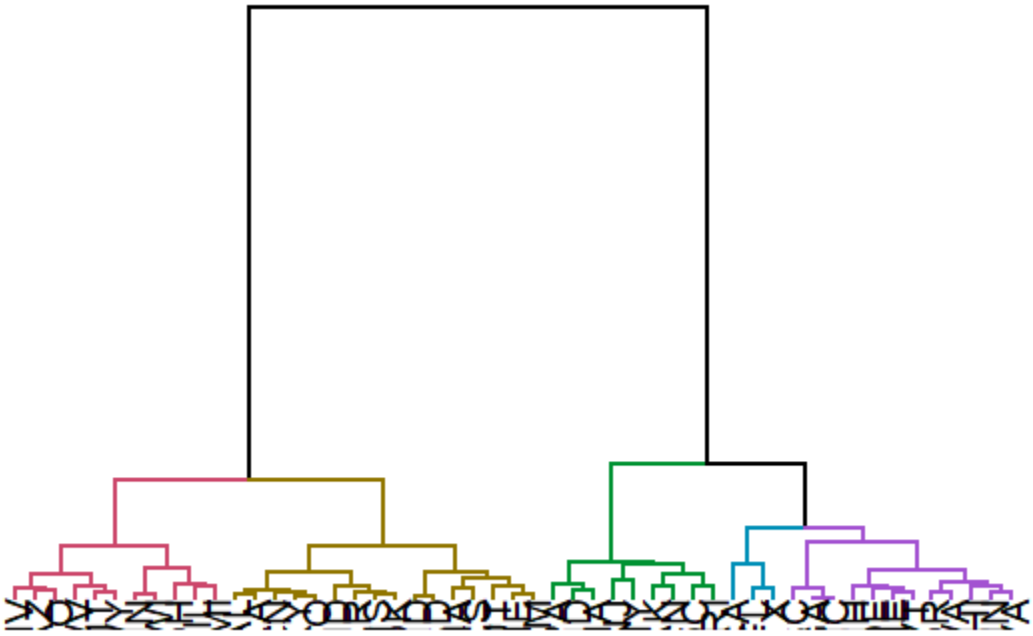
```
#cortar el dendograma en 5 grupos
clust <- cutree(hc1, k = 5)

#Gráfico de pertenencia de individuos al clúster
fviz_cluster(list(data = datos2, cluster = clust))
```



11. Realiza las conclusiones de todo el análisis del clúster jerárquico.

Podemos concluir que este análisis nos permite identificar grupos en el dataset y nos ayuda a comprender la estructura de estos mismos así como a interpretar la similitud que hay entre el conjunto de datos.





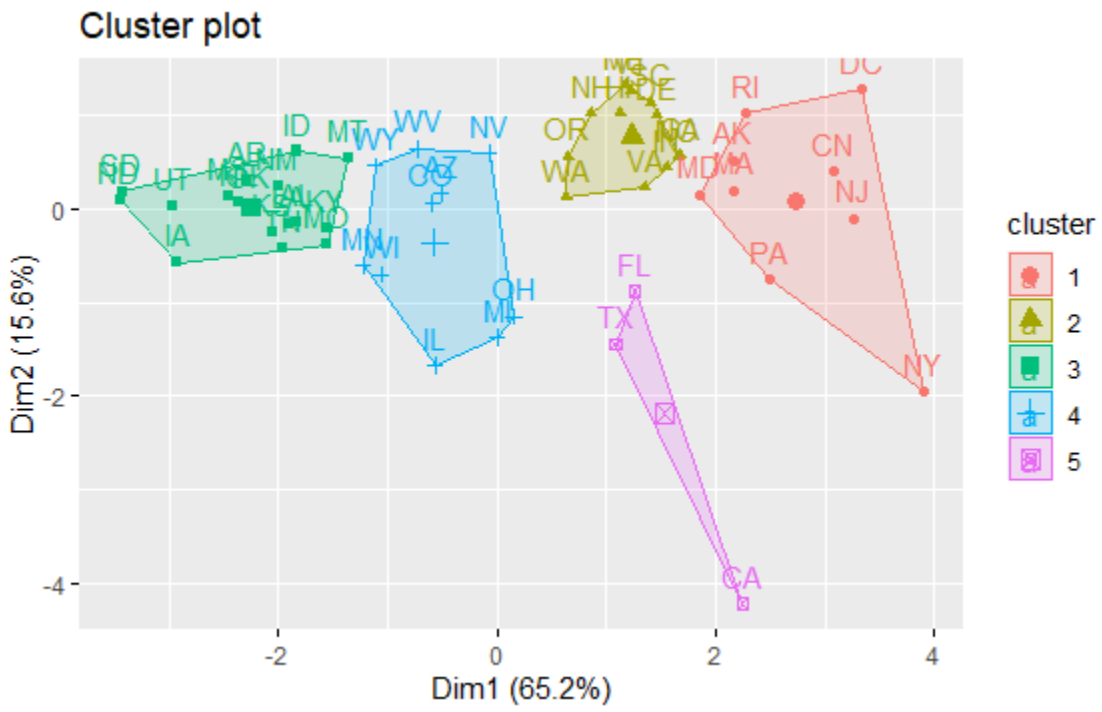
12. Realiza el análisis de clúster con el método Kmeans.

Hemos especificado 5 grupos para este algoritmo ya que con estos grupos el clustering explica el 76.3% de la variabilidad total en los datos, por otro lado pensamos en tomar 4 grupos pero descartamos esa opción ya que se explicaba menos del 70%.

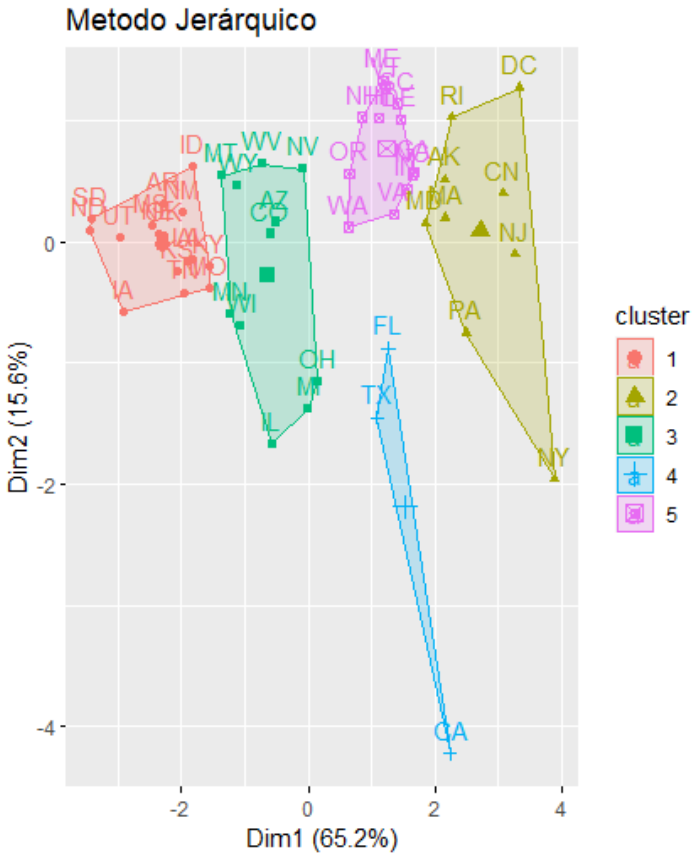
```
> kmedias<-kmeans(datos2, 5)
> kmedias
K-means clustering with 5 clusters of sizes 9, 12, 17, 10, 3

Cluster means:
          pop       SATV       SATM    percent     dollars         pay
1  0.20338309 -0.8198294 -0.8695358  1.2752117  1.70752892  1.53912994
2 -0.27289178 -0.9108568 -0.9249807  0.9175171  0.04796646  0.01108174
3 -0.42913945  1.1095053  0.9824109 -1.0107756 -0.85259611 -0.96411099
4  0.04562304  0.2512292  0.4572862 -0.6540313 -0.04984152  0.16179334
5  2.76113133 -1.0217119 -0.7827524  0.4121293 -0.31693624  0.26226774

Clustering vector:
AL AK AZ AR CA CO CN DE DC FL GA HI ID IL IN IA KS KY LA ME MD MA MI MN MS MO MT NE NV NH
 3  1  4  3  5  4  1  2  1  5  2  2  3  4  2  3  3  3  3  2  1  1  4  4  3  3  3  3  4  2
NJ NM NY NC ND OH OK OR PA RI SC SD TN TX UT VT VA WA WV WI WY
 1  3  1  2  3  4  3  2  1  1  2  3  3  5  3  2  2  2  4  4  4

Within cluster sum of squares by cluster:
[1] 19.102822 12.269660 16.850829 14.797098  8.150464
 (between_SS / total_SS =  76.3 %)

Available components:

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

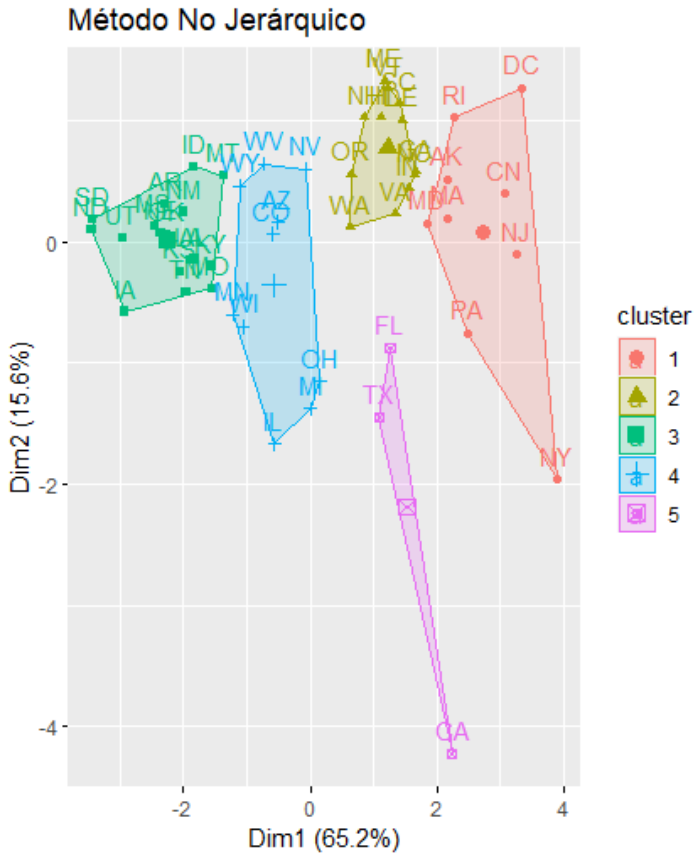
Podemos observar que se ha realizado un análisis de clustering k-means con un total de 5 clústeres, en el recuadro verde podemos observar cuántas observaciones pertenecen a cada uno de los clústeres. El primer clúster tiene 9 observaciones, el segundo tiene 12, el tercero tiene 17, el cuarto tiene 10 y el último tiene 3 .

En el recuadro amarillo tenemos las medias de cada variable para cada uno de los 5 clústeres. El clustering vector muestra a qué clúster pertenece cada observación y en el recuadro rojo el clustering explica aproximadamente el 76.3% de la variabilidad total en los datos.



13. Compara los resultados del método jerárquico y el no jerárquico.





La observación de que en el método jerárquico el cluster número "1" estaba en la izquierda del dendrograma, mientras que en el método no jerárquico estaba en la derecha, así mismo pasa con el resto, puede tener implicaciones interesantes en la interpretación y elección del método de agrupamiento adecuado para los datos en cuestión.

Una de las causas pudo ser:
-Sensibilidad a la estructura jerárquica: La diferencia en la posición del cluster número "1" en el dendrograma sugiere que el método jerárquico puede ser más sensible a la estructura jerárquica intrínseca en los datos. Esto podría ser ventajoso si se busca una comprensión más profunda de cómo se organizan los datos en múltiples niveles de similitud.

NOTA: debes subir a SAVI dos archivos

1. **El archivo R**
2. **El Word con las explicaciones**