**Ejercicio 1**

Dado el conjunto de datos representado por la matriz:

i. Grafique en R2 y construya el dendrograma correspondiente utilizando el criterio del

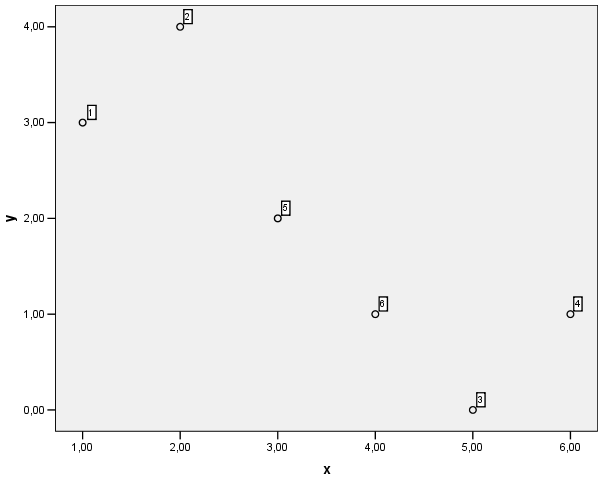
vecino más lejano (utilizar la distancia euclídea).

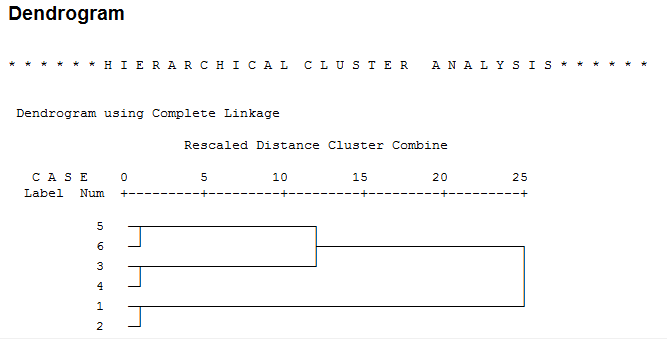
ii. Igual que el anterior, utilizando el criterio de vecino más cercano.

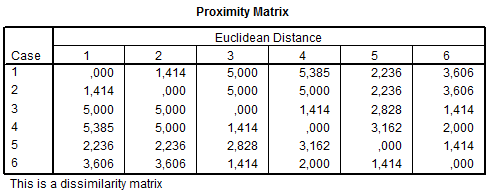
iii. Repita el item 1.i. pero aplicando el criterio promedio

iv. Repita el ejercicio utilizando las variables estandarizadas. Compare los resultados.

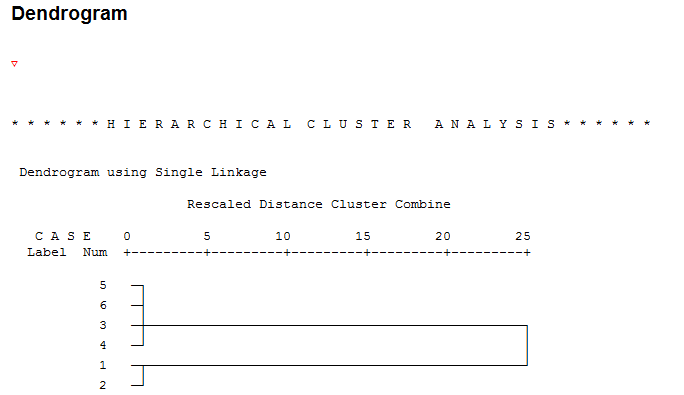
1. Vecino más lejano (Distancia Euclídea) - Complete Linkage

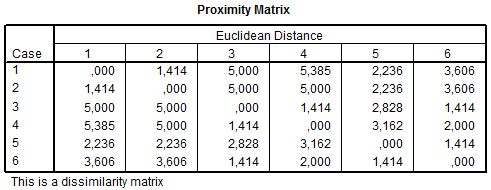




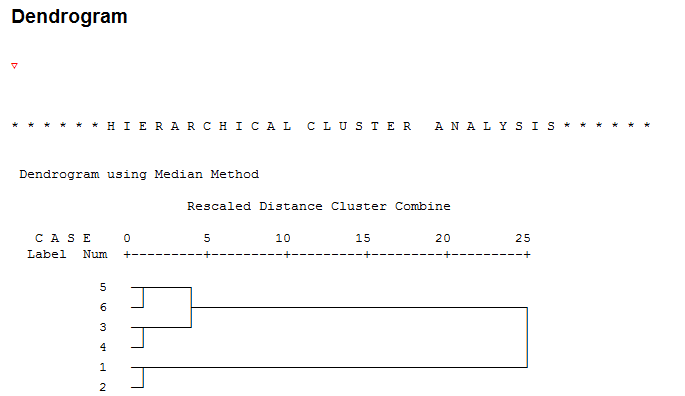


1. Vecino más cercano (Distancia Euclídea) – Single Linkage

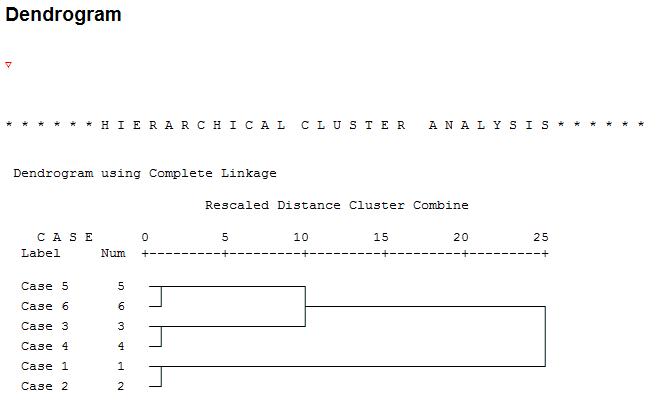


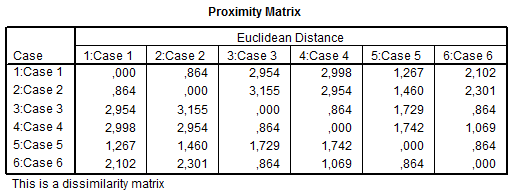


1. Media (Distancia Euclídea) – Average Linkage

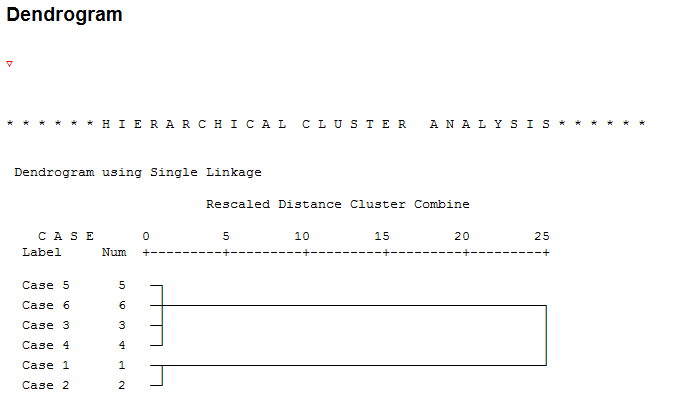


1. Vecino más lejano





Vecino más cercano



Media

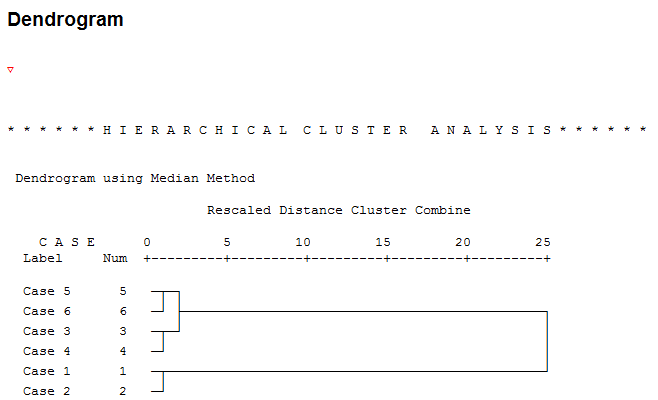
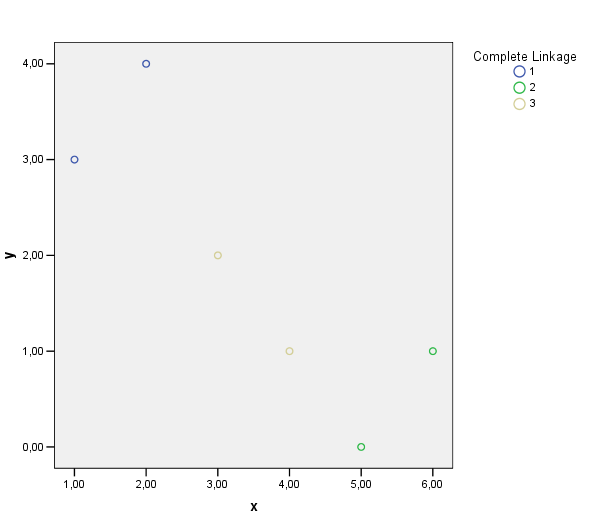


Gráfico del ejercicio con 3 cluster:



**Ejercicio 2**

Dada la siguiente matriz de distancias, realice los dendrogramas correspondientes a los métodos: vecino más cercano, vecino más lejano y promedio. (distancia euclídea)

¿Encuentra diferencias entre los resultados obtenidos?

**Ejercicio 3:** Pizzas

Si se desea obtener cinco agrupamientos de los datos correspondientes a la tabla ‘**pizzas**’:

i. Realice un Análisis en Componentes Principales. ¿Qué proporción de la variabilidad

total en las variables medidas explican las dos primeras componentes? Utilizando un gráfico de individuos determinar grupos en los datos. ¿Cuántos grupos hay? ¿Cuáles

pizzas pertenecen a cuáles agrupamientos? Comparar con el ítem anterior.

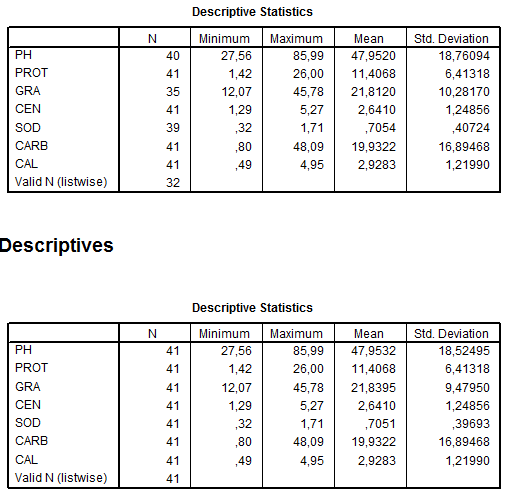
ii. Aplique un método de agrupamiento a los resultados del ítem anterior (valores de los casos sobre las componentes).

iii. Aplique el método de K-Medias a los datos de manera de obtener 5 grupos. Compare con los resultados anteriores.

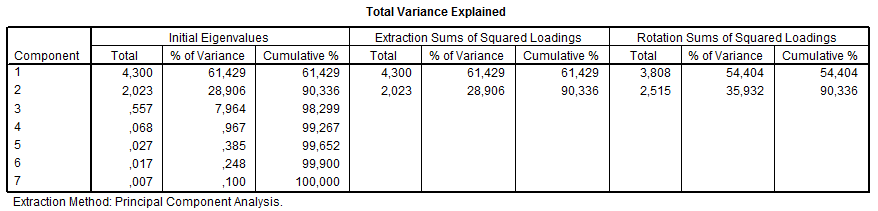
iv. Resuma los resultados: ¿tienen los datos una estructura como para agruparlos? En el caso de que su respuesta sea afirmativa: ¿en cuántos grupos le parece más conveniente?

Justifique.

1. 1) Mejora de los datos

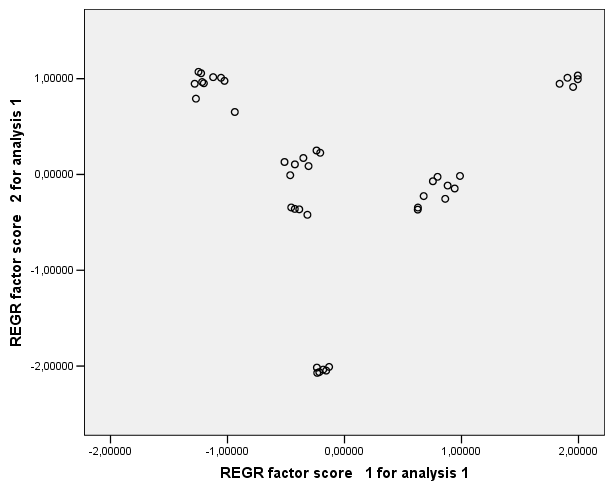


2) Componentes Principales:

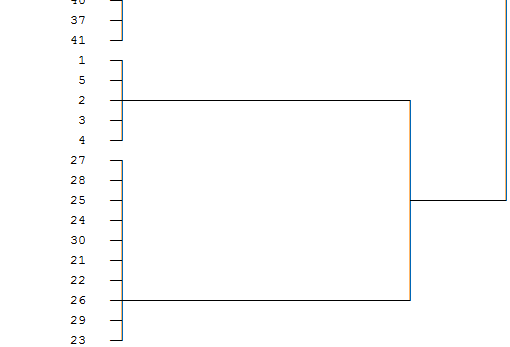
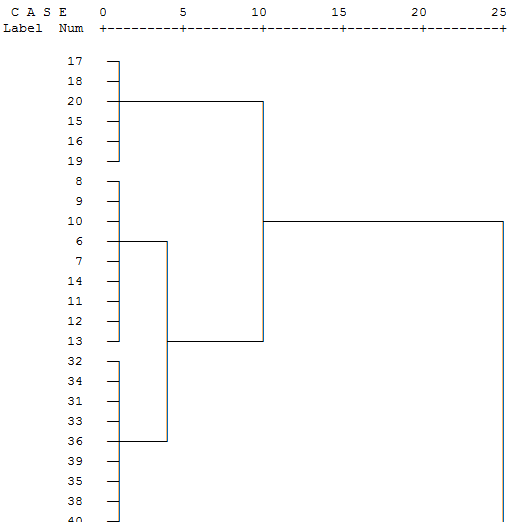


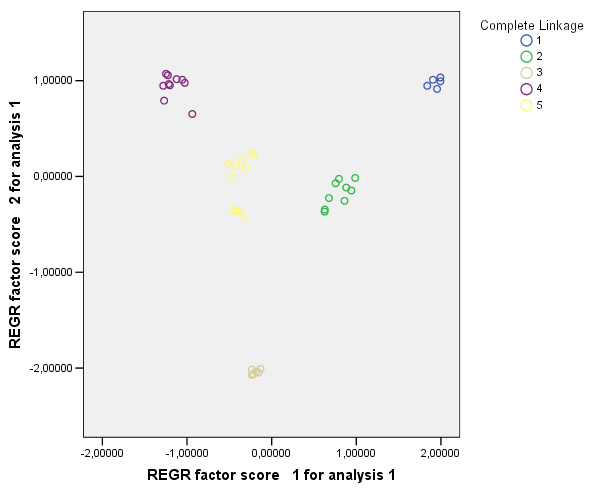
Con la reducción a dos componentes se explica el 90.34 %

3) Grafico en R2: Se observan 5 grupos posibles graficando las dos componentes

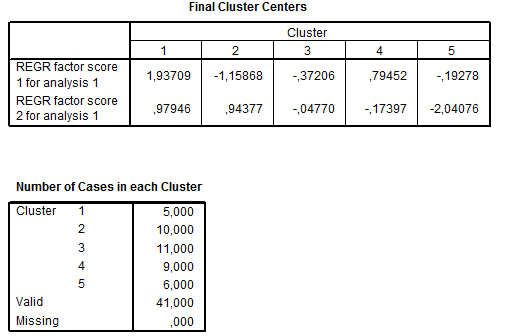


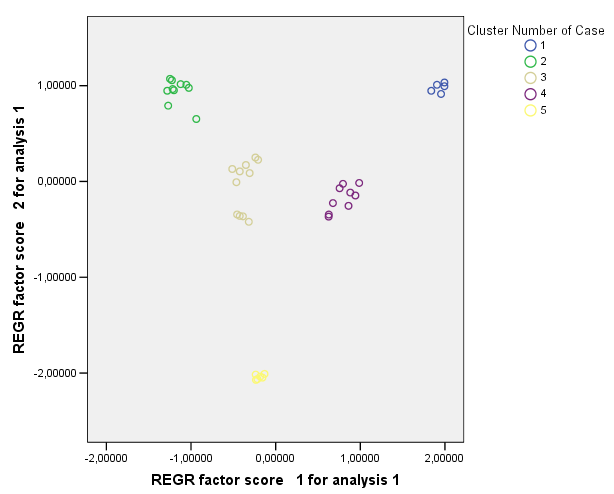
1. Dendrograma





iii)

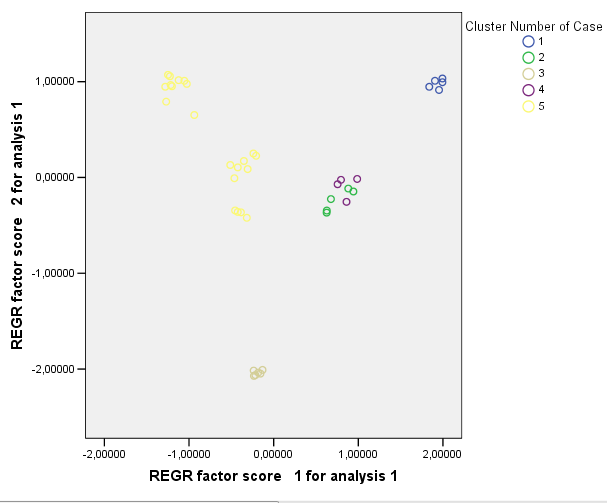




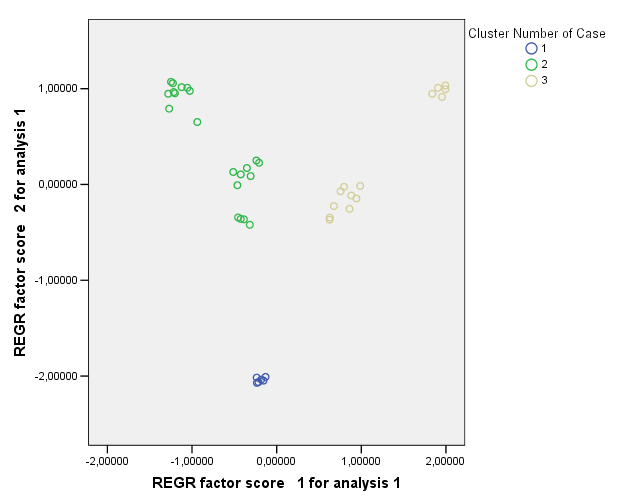
Los resultados son idénticos a los anteriores.

iv)Los datos parecerían tener estructura para ser agrupados en 5 grupos. Esto se pudo visualizar con los dos componentes que explicaban el 90%. Realizaremos un agrupamiento sobre los datos originales estandarizados para ver que ocurre realmente:

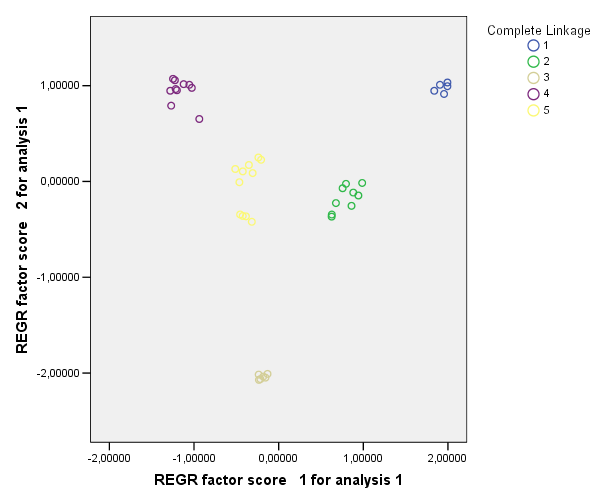
Visualizando el resultado de los 5 grupos (k means) sobre los 2 ejes de las componentes, no parece separarse tan bien los 5 grupos:



En 3 cluster (k means) el agrupamiento parece visualizarse mejor:



Probemos Jerárquicos:



Con vecino más lejano (distancia euclídea2) los 5 agrupamientos se pueden visualizar con los 2 componentes que explicaban el 90%.

**Por lo tanto, la conclusión final es que si conviene agrupar en 5 tipos de pizzas.**

**Ejercicio 4:** Museos

Supongamos que un Museo realiza encuestas a un grupo de niños al terminar el recorrido; dicha encuesta ésta diseñada con distintas preguntas generales y algunas que pueden ayudarnos a identificar grupos y diseñar estrategias que vayan acorde con los niños que están más interesados en asistir a un museo.

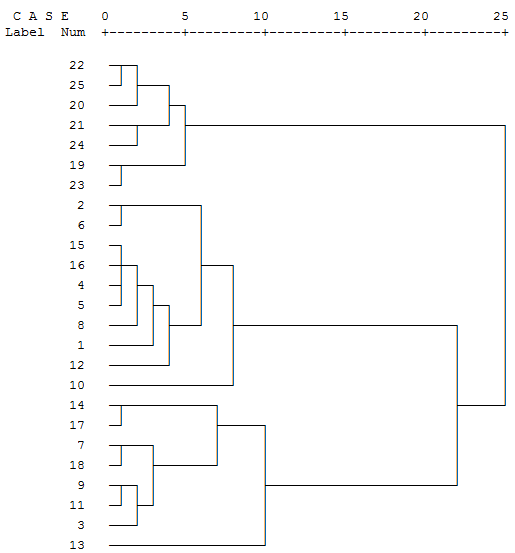
Algunas de las preguntas que encontramos en ésta encuesta son las siguientes:

|  |
| --- |
| Sexo - Edad |
| A ¿Es divertido ir al museo?**\* *divertid*** |
| B .Cuando voy al museo le pido a mis papas que me compren algo de lo que venden adentro? **\* *pidocomp*** |
| C ¿Puedo aprender en la escuela lo mismo que en el museo? **\* *aprendom*** |
| D ¿Prefiero ir al museo en excursiones con la escuela? **\*  *excur*** |
| E ¿Ir al museo en mi tiempo libre me quita tiempo para jugar? **\* *quitatie*** |
| F ¿No me interesa en lo más mínimo asistir al museo? **\* *nomeint*** |
| ¿Te gustó tu visita al museo? (Si/No) ***gustovis*** |

***\**** ***De Totalmente en desacuerdo a Totalmente de acuerdo en escala de 7 puntos.***

Suponiendo que las preguntas de la sección denominada como “Opiniones generales que tengo en relación al museo” son con las que el equipo pretende agrupar a los 25 niños que respondieron la encuesta, se tendrían que hacer primero 2 consideraciones:

1. Que para identificar los grupos de niños, las preguntas que se elijan deben de estar en la misma escala de medición. (En caso de que esto no sea posible, se deben de estandarizar los valores)
2. Justificar la aplicación del cluster análisis, lo cual podemos hacer demostrando que existe fuerte asociación entre las variables que van a configurarlo.

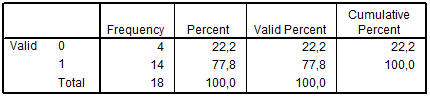


|  |
| --- |
| Grupo 1 |
|  |

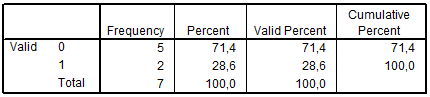
|  |
| --- |
| Grupo 2 |
|  |

Frecuencia de Si le gusto o no la visita el museo:

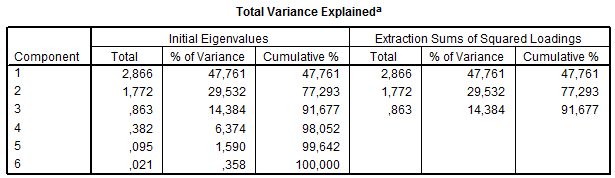
Grupo 1

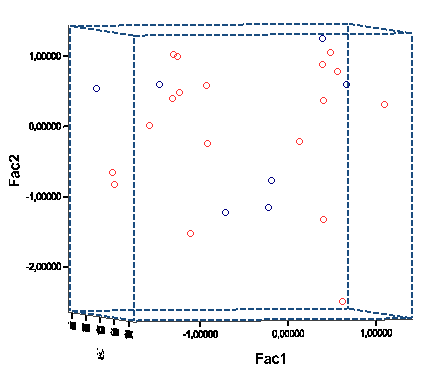


Grupo 2



Componentes Principales





**Ejercicio 5: Alumnos de psicología**

Queremos agrupar a 6 alumnos de primero de psicología en base a sus notas en las asignaturas del área de básica (X1), del área de metodología (X2), del área de evolutiva (X3), del área de social (X4) y del área de clínica (X5). Para ello hemos realizado la media por área y hemos obtenido la siguiente matriz:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 |
| S1 | 8 | 9 | 7 | 8 | 6 |
| S2 | 7 | 8 | 7 | 8 | 8 |
| S3 | 2 | 3 | 8 | 7 | 2 |
| S4 | 1 | 2 | 6 | 7 | 1 |
| S5 | 1 | 1 | 1 | 9 | 8 |
| S6 | 2 | 3 | 1 | 8 | 9 |

Con los datos de la matriz anterior realizar los dendogramas, utilizando el método de Ward:

1. Para los datos crudos
2. Para los datos estandarizados por variable
3. ¿A qué se deben las diferencias observadas en los dendrogramas?
4. Cuál de las alternativas seleccionaría teniendo en cuenta este coeficiente y la interpretabilidad de los resultados.

a y b)

|  |  |
| --- | --- |
| Variables Sin estandarizar | Variables estandarizadas |
|  |  |

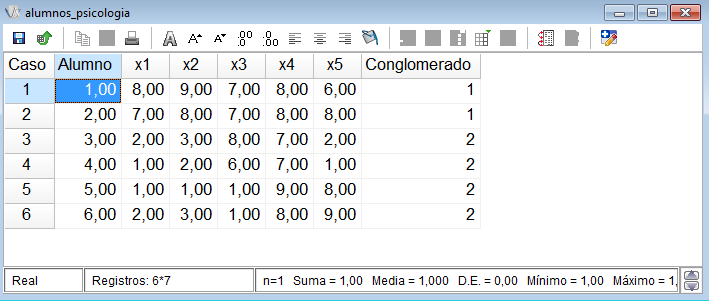
|  |  |
| --- | --- |
| Variables Sin estandarizar | Variables estandarizadas |
|  |  |

c) Las diferencias observadas son: En el caso de las variables originales sin estandarizar, si se definen dos cluster, el tercer cluster (observaciones 5 y 6) se unifica con el cluster que contienen las observaciones (3 y 4). Pero si se toman las variables estandarizadas el tercer cluster (observaciones 5 y 6) se unifica con el cluster (1 2).

Evidentemente ante la necesidad de seleccionar 2 cluster solamente el resultado sería totalmente distinto en utilizar las variables estandarizadas o no, porque una parte de la población se inclinaría hacia un lado u hacia el otro.

d)

***En este caso no es necesario estandarizar las variables. Porque las notas son valores entre 1 y 10. Y todas las variables tienen el mismo peso entre sí. Por lo tanto en este caso conviene tomar las variables reales sin estandarizar.***



**Ejercicio 6:** Consumo de proteínas en varios países Europeos.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| País | Carne  Vacuna | Carne  Cerdo | Huevos | Leche | Pescado | Cereal | Embutidos | Frutos  Secos | Frutas y Vegetales |
| Albania | 10.1 | 1.4 | 0.5 | 8.9 | 0.2 | 42.3 | 0.6 | 5.5 | 1.7 |
| Austria | 8.9 | 14 | 4.3 | 19.9 | 2.1 | 28 | 3.6 | 1.3 | 4.3 |
| Bélgica | 13.5 | 9.3 | 4.1 | 17.5 | 4.5 | 26.6 | 5.7 | 2.1 | 4 |
| Bulgaria | 7.8 | 6 | 1.6 | 8.3 | 1.2 | 56.7 | 1.1 | 3.7 | 4.2 |
| Checosl | 9.7 | 11.4 | 2.8 | 12.5 | 2 | 34.3 | 5 | 1.1 | 4 |
| Dinamarca | 10.6 | 10.8 | 3.7 | 25 | 9.9 | 21.9 | 4.8 | 0.7 | 2.4 |
| AlemaniaE | 8.4 | 11.6 | 3.7 | 11.1 | 5.4 | 24.6 | 6.5 | 0.8 | 3.6 |
| Finlandia | 9.5 | 4.9 | 2.7 | 33.7 | 5.8 | 26.3 | 5.1 | 1 | 1.4 |
| Francia | 18 | 9.9 | 3.3 | 19.5 | 5.7 | 28.1 | 4.8 | 2.4 | 6.5 |
| Grecia | 10.2 | 3 | 2.8 | 17.6 | 5.9 | 41.7 | 2.2 | 7.8 | 6.5 |
| Hungría | 5.3 | 12.4 | 2.9 | 9.7 | 0.3 | 40.1 | 4 | 5.4 | 4.2 |
| Irlanda | 13.9 | 10 | 4.7 | 25.8 | 2.2 | 24 | 6.2 | 1.6 | 2.9 |
| Italia | 9 | 5.1 | 2.9 | 13.7 | 3.4 | 36.8 | 2.1 | 4.3 | 6.7 |
| P.Bajos | 9.5 | 13.6 | 3.6 | 23.4 | 2.5 | 22.4 | 4.2 | 1.8 | 3.7 |
| Noruega | 9.4 | 4.7 | 2.7 | 23.3 | 9.7 | 23 | 4.6 | 1.6 | 2.7 |
| Polonia | 6.9 | 10.2 | 2.7 | 19.3 | 3 | 36.1 | 5.9 | 2 | 6.6 |
| Portugal | 6.2 | 3.7 | 1.1 | 4.9 | 14.2 | 27 | 5.9 | 4.7 | 7.9 |
| Rumania | 6.2 | 6.3 | 1.5 | 11.1 | 1 | 49.6 | 3.1 | 5.3 | 2.8 |
| España | 7.1 | 3.4 | 3.1 | 8.6 | 7 | 29.2 | 5.7 | 5.9 | 7.2 |
| Suecia | 9.9 | 7.8 | 3.5 | 24.7 | 7.5 | 19.5 | 3.7 | 1.4 | 2 |
| Suiza | 13.1 | 10.1 | 3.1 | 23.8 | 2.3 | 25.6 | 2.8 | 2.4 | 4.9 |
| Inglaterra | 17.4 | 5.7 | 4.7 | 20.6 | 4.3 | 24.3 | 4.7 | 3.4 | 3.3 |
| Rusia | 9.3 | 4.6 | 2.1 | 16.6 | 3 | 43.6 | 6.4 | 3.4 | 2.9 |
| AlemaniaO | 11.4 | 12.5 | 4.1 | 18.8 | 3.4 | 18.6 | 5.2 | 1.5 | 3.8 |

1. Utilizando el método de Ward y la distancia euclidea particionar en dos clusters. Como llamaría a cada uno de ellos?
2. Idem a- pero en cuatro clusters. Utilizando el dendograma, con cuál de las clasificaciones se quedaría?
3. Realice una cauterización de las variables?.
4. Compare los resultados obtenidos con el de componentes principales.