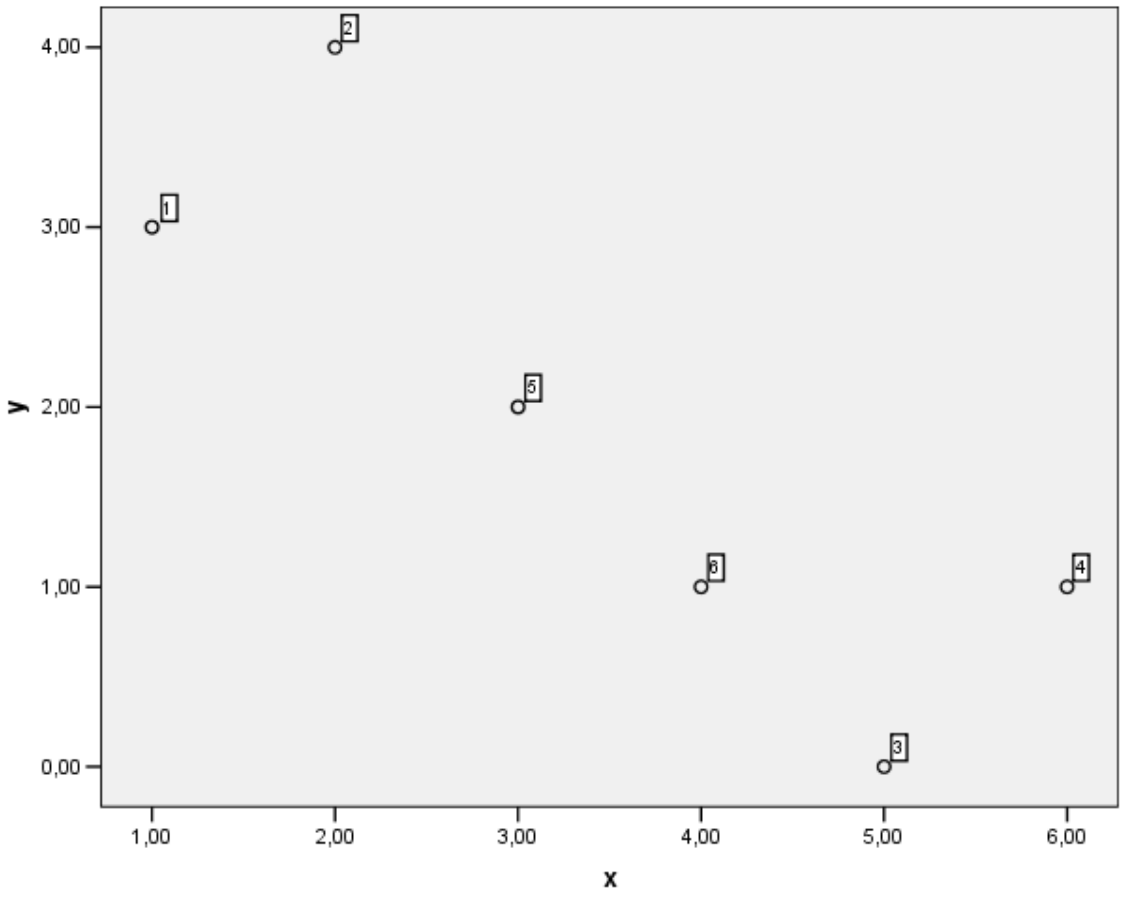
Ejercicio 1

Dado el conjunto de datos representado por la matriz:

$$X = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 5 & 0 \\ 6 & 1 \\ 3 & 2 \\ 4 & 1 \end{pmatrix}$$

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.

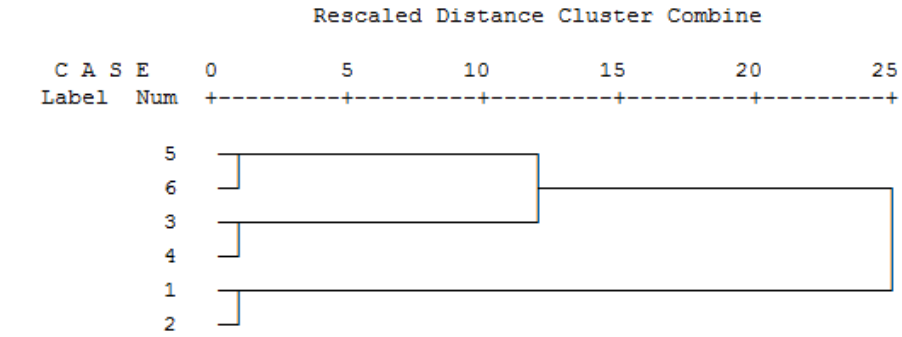
i.) Vecino más lejano (Distancia Euclídea) - Complete Linkage



Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Complete Linkage



Proximity Matrix

	Euclidean Distance					
Case	1	2	3	4	5	6
1	,000	1,414	5,000	5,385	2,236	3,606
2	1,414	,000	5,000	5,000	2,236	3,606
3	5,000	5,000	,000	1,414	2,828	1,414
4	5,385	5,000	1,414	,000	3,162	2,000
5	2,236	2,236	2,828	3,162	,000	1,414
6	3,606	3,606	1,414	2,000	1,414	,000

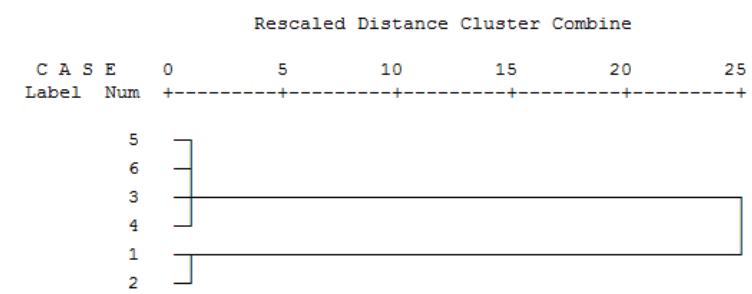
This is a dissimilarity matrix

ii) Vecino más cercano (Distancia Euclídea) – Single Linkage

Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Single Linkage



Proximity Matrix

Case	Euclidean Distance					
	1	2	3	4	5	6
1	,000	1,414	5,000	5,385	2,236	3,606
2	1,414	,000	5,000	5,000	2,236	3,606
3	5,000	5,000	,000	1,414	2,828	1,414
4	5,385	5,000	1,414	,000	3,162	2,000
5	2,236	2,236	2,828	3,162	,000	1,414
6	3,606	3,606	1,414	2,000	1,414	,000

This is a dissimilarity matrix

iii) Media (Distancia Euclídea) – Average Linkage

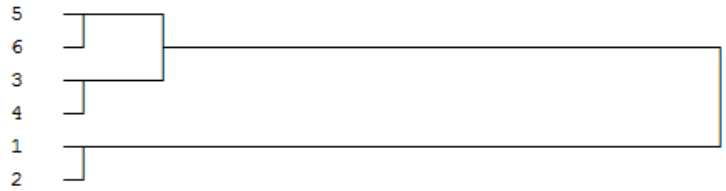
Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Median Method

Rescaled Distance Cluster Combine

```
C A S E    0         5         10        15        20        25
Label  Num +---------+---------+---------+---------+---------+
```



iv) Vecino más lejano

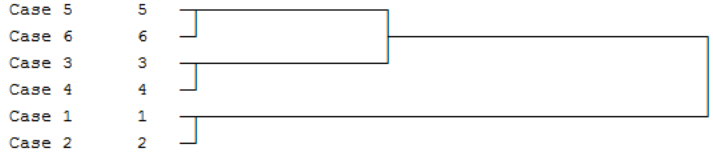
Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Complete Linkage

Rescaled Distance Cluster Combine

```
C A S E      0         5         10        15        20        25
Label    Num +---------+---------+---------+---------+---------+
```



Proximity Matrix

Case	Euclidean Distance					
	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6
1:Case 1	,000	,864	2,954	2,998	1,267	2,102
2:Case 2	,864	,000	3,155	2,954	1,460	2,301
3:Case 3	2,954	3,155	,000	,864	1,729	,864
4:Case 4	2,998	2,954	,864	,000	1,742	1,069
5:Case 5	1,267	1,460	1,729	1,742	,000	,864
6:Case 6	2,102	2,301	,864	1,069	,864	,000

This is a dissimilarity matrix

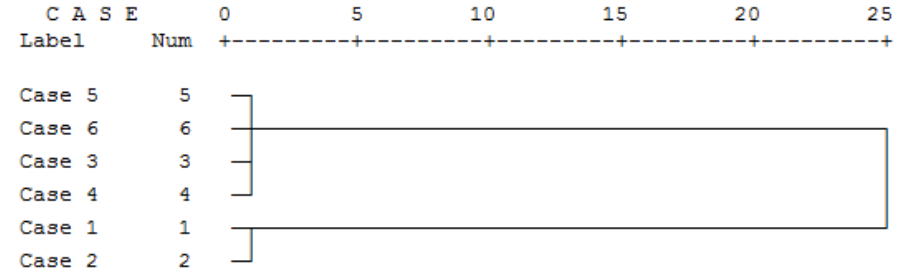
Vecino más cercano

Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Single Linkage

Rescaled Distance Cluster Combine



Media

Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Median Method

Rescaled Distance Cluster Combine

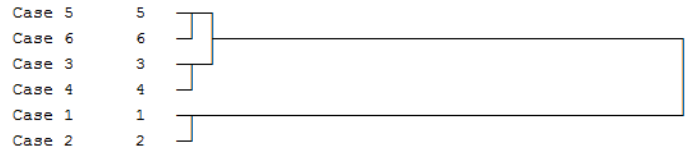
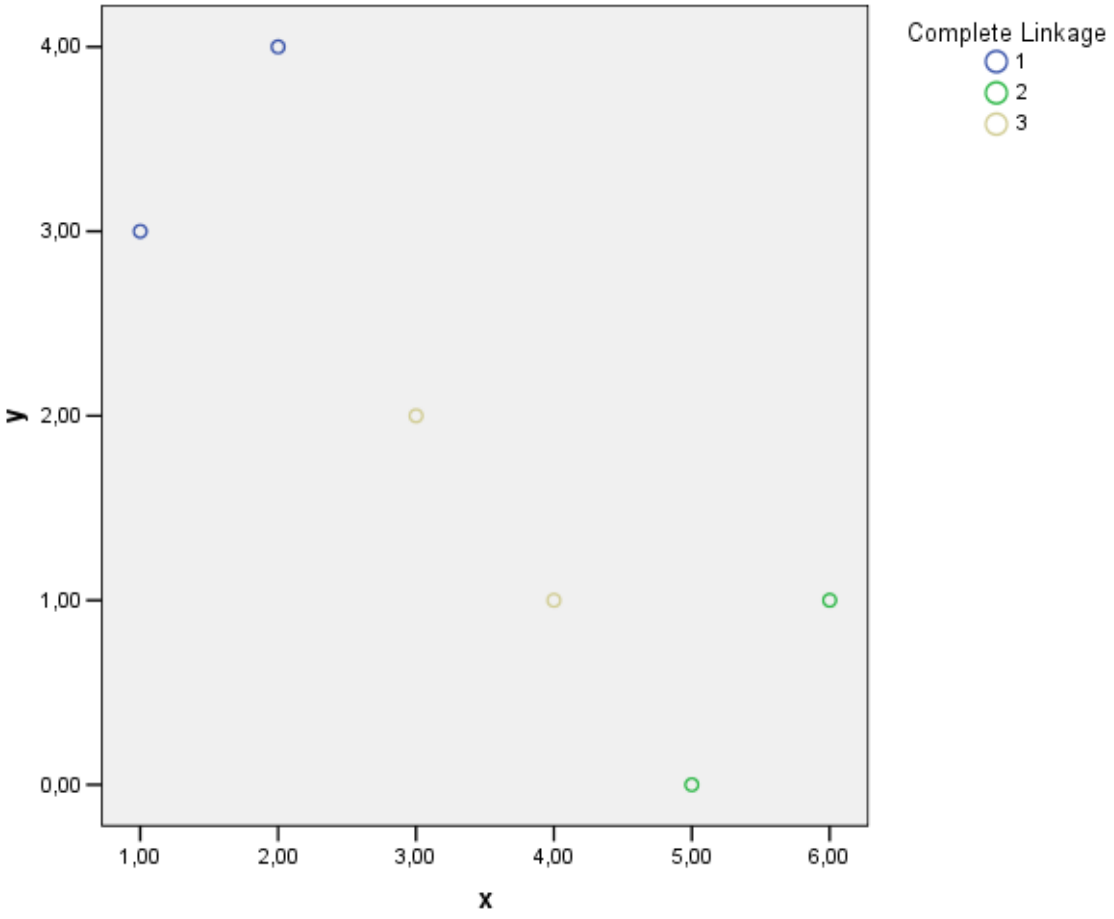


Gráfico del ejercicio con 3 cluster:



Ejercicio 2

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

$$D = \begin{pmatrix} 0 & & & & \\ 4 & 0 & & & \\ 18 & 10 & 0 & & \\ 20 & 15 & 24 & 0 & \\ 18 & 20 & 8 & 6 & 0 \end{pmatrix}$$

¿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.

i) 1) Mejora de los datos

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PH	40	27,56	85,99	47,9520	18,76094
PROT	41	1,42	26,00	11,4068	6,41318
GRA	35	12,07	45,78	21,8120	10,28170
CEN	41	1,29	5,27	2,6410	1,24856
SOD	39	,32	1,71	,7054	,40724
CARB	41	,80	48,09	19,9322	16,89468
CAL	41	,49	4,95	2,9283	1,21990
Valid N (listwise)	32				

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PH	41	27,56	85,99	47,9532	18,52495
PROT	41	1,42	26,00	11,4068	6,41318
GRA	41	12,07	45,78	21,8395	9,47950
CEN	41	1,29	5,27	2,6410	1,24856
SOD	41	,32	1,71	,7051	,39693
CARB	41	,80	48,09	19,9322	16,89468
CAL	41	,49	4,95	2,9283	1,21990
Valid N (listwise)	41				

2) Componentes Principales:

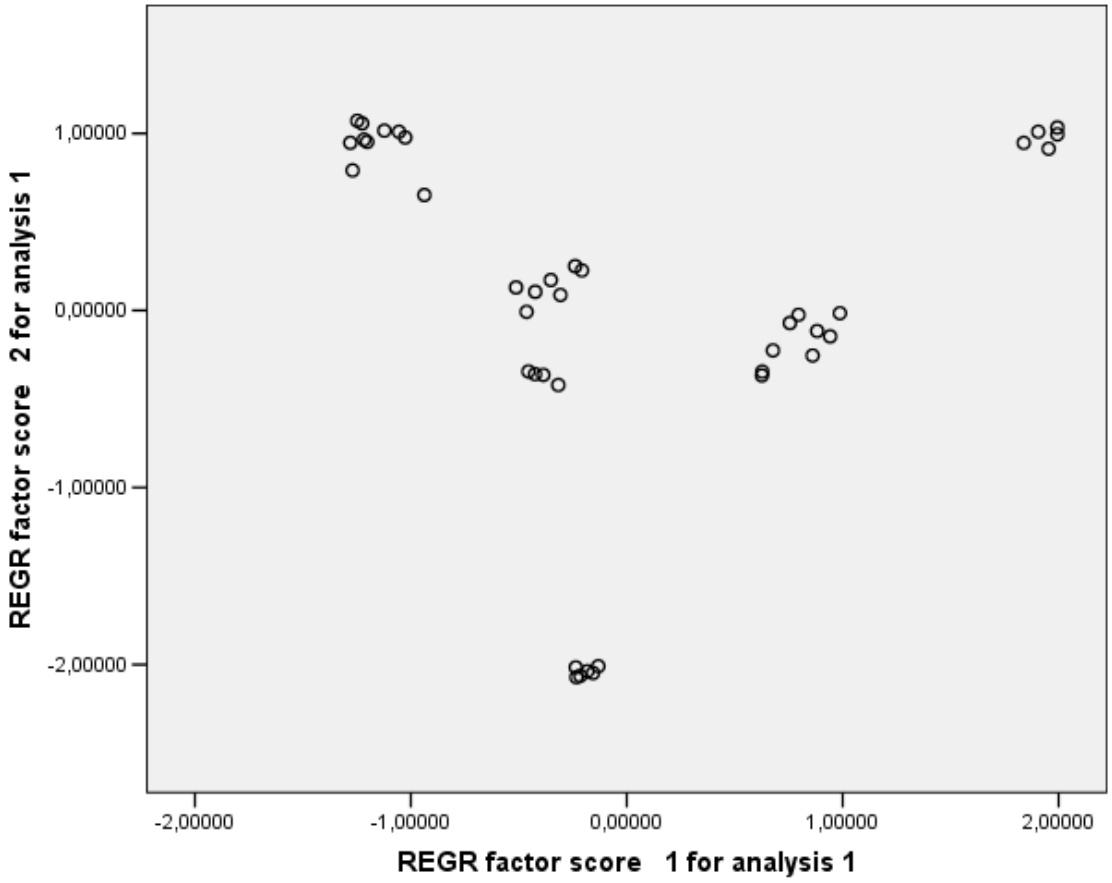
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,300	61,429	61,429	4,300	61,429	61,429	3,808	54,404	54,404
2	2,023	28,906	90,336	2,023	28,906	90,336	2,515	35,932	90,336
3	,557	7,964	98,299						
4	,068	,967	99,267						
5	,027	,385	99,652						
6	,017	,248	99,900						
7	,007	,100	100,000						

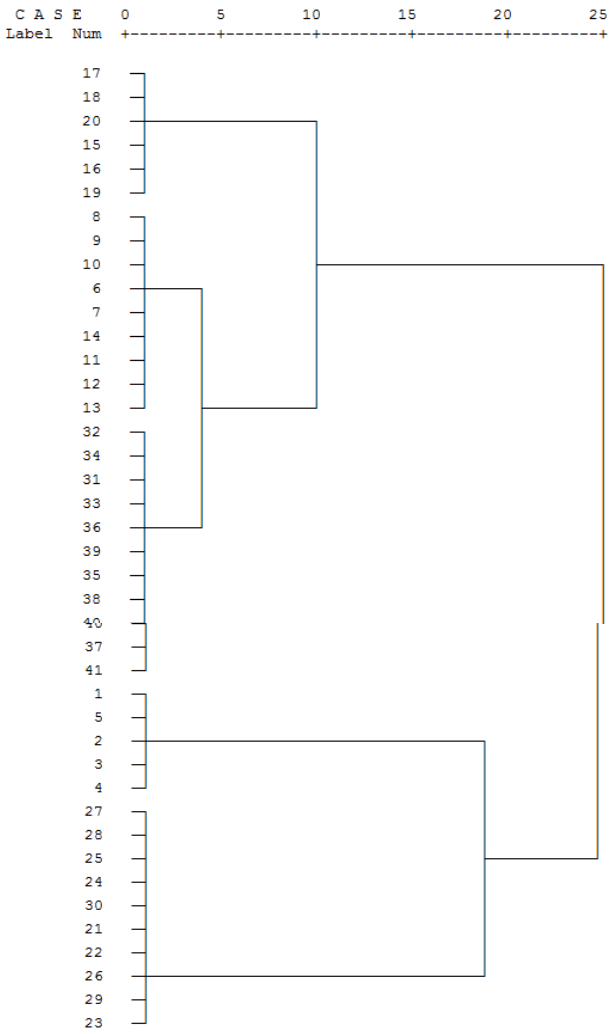
Extraction Method: Principal Component Analysis.

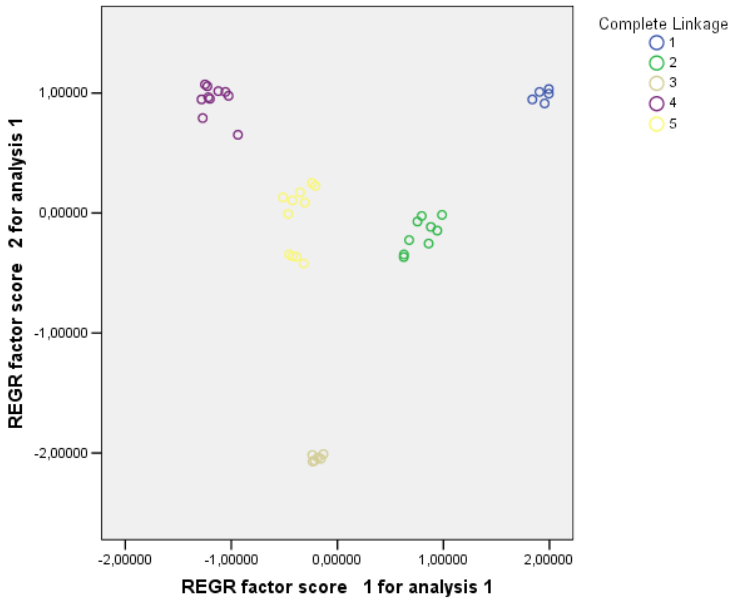
Con la reducción a dos componentes se explica el 90.34 %

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



ii) Dendrograma





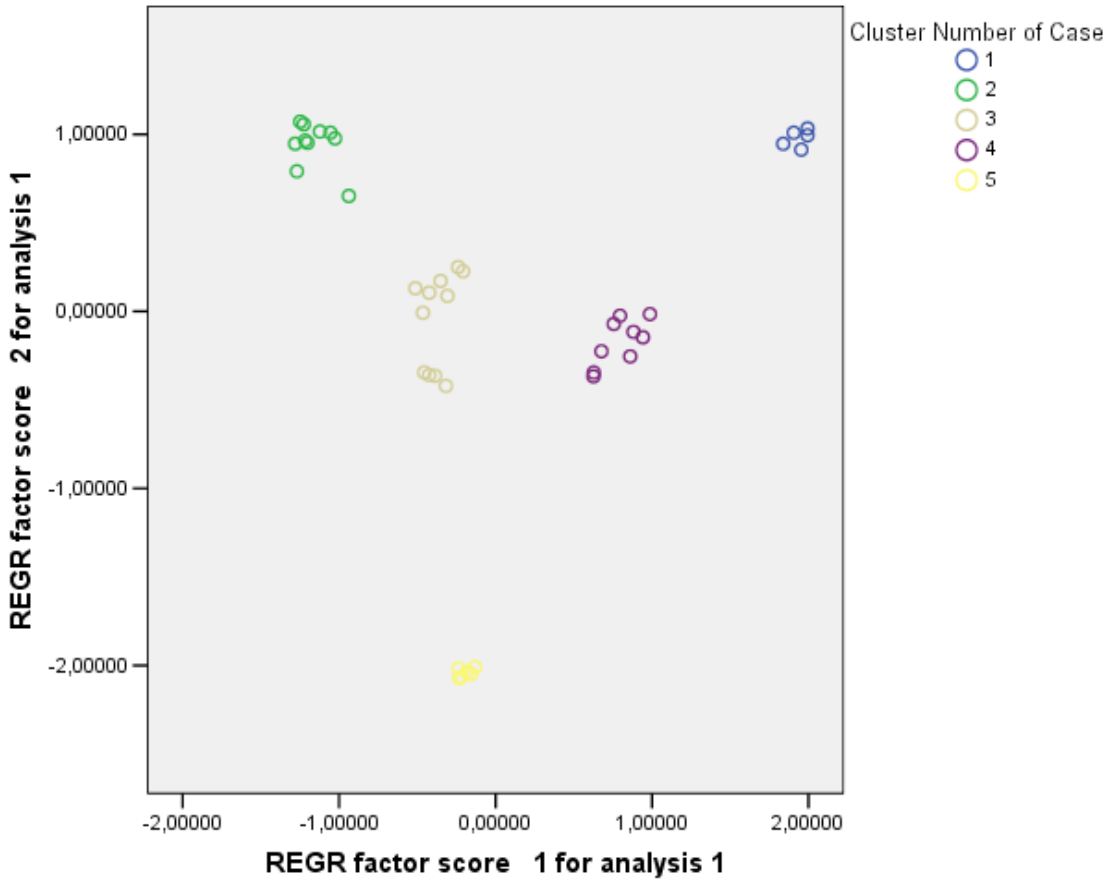
iii)

Final Cluster Centers

	Cluster				
	1	2	3	4	5
REGR factor score 1 for analysis 1	1,93709	-1,15868	-,37206	,79452	-,19278
REGR factor score 2 for analysis 1	,97946	,94377	-,04770	-,17397	-2,04076

Number of Cases in each Cluster

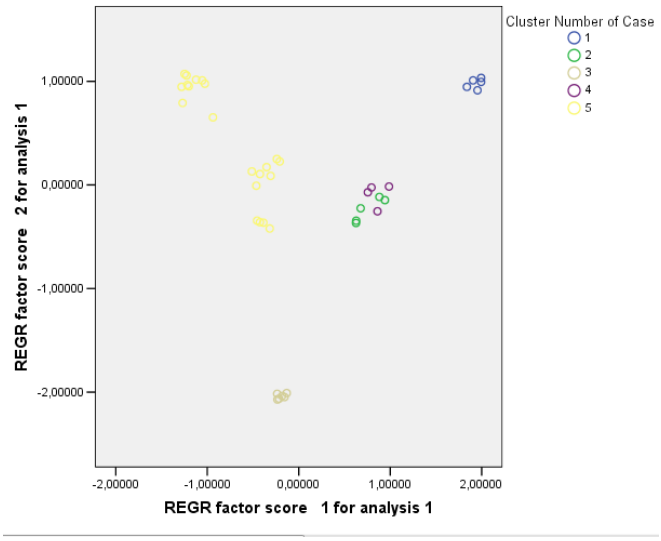
Cluster	1	5,000
	2	10,000
	3	11,000
	4	9,000
	5	6,000
Valid		41,000
Missing		,000



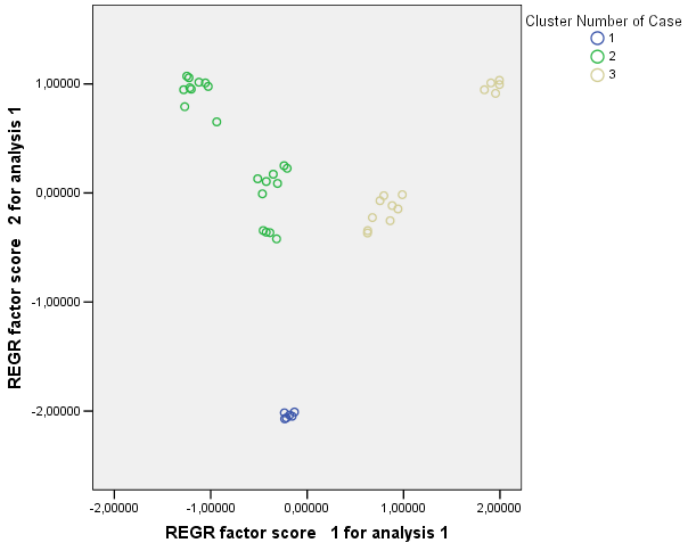
Los resultados son idénticos a los anteriores.

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:

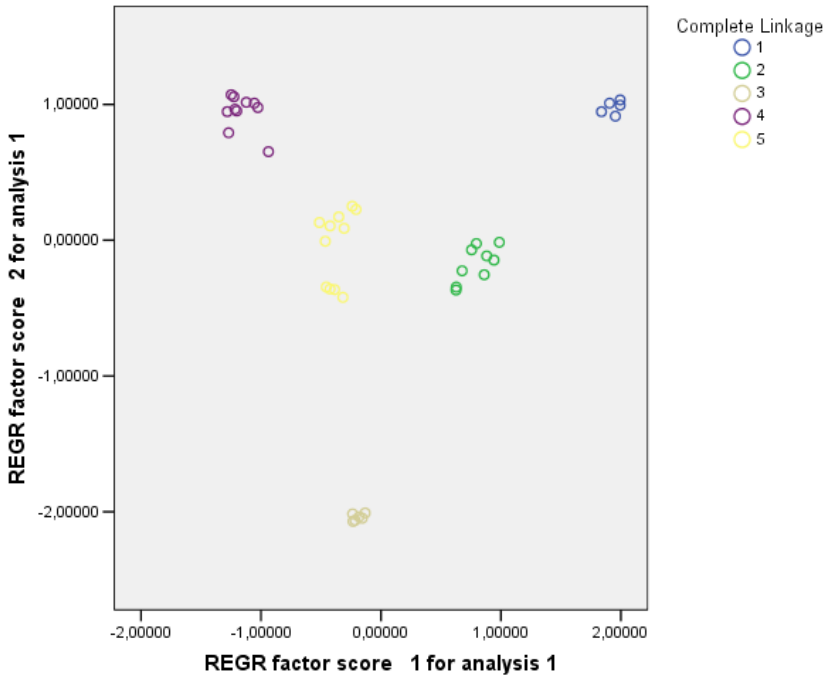
Visualizando el resultado de los 5 grupos (k means) sobre los 2 ejes de las componentes, no parece separarse tan bien los 5 grupos:



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

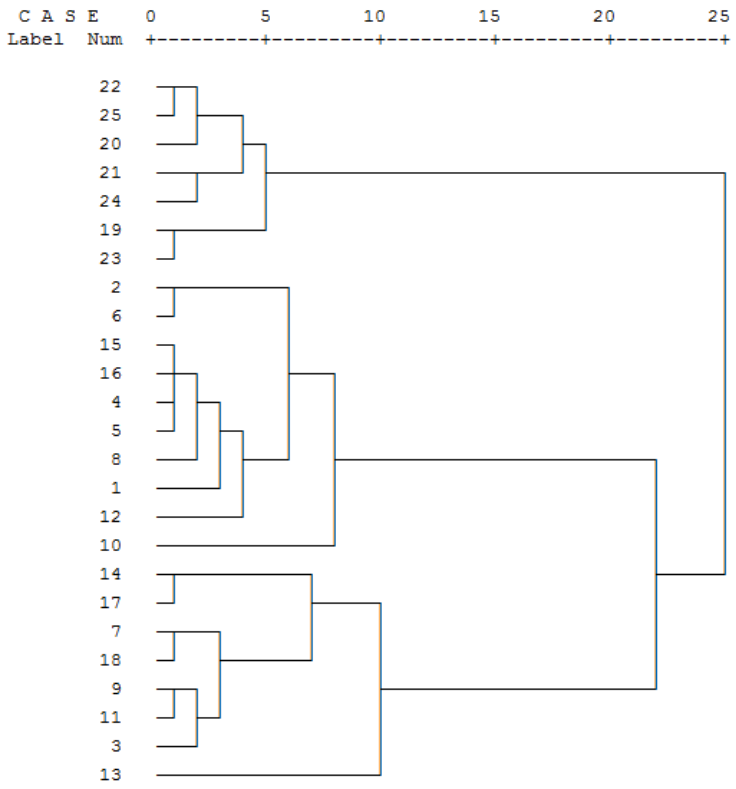
¿Te gustó tu visita al museo? (Si/No)

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

	N	Minimum	Maximum	Mean	Std. Deviation
divertid	18	3	7	4,67	1,328
pidocomp	18	2	7	4,78	1,768
aprendom	18	3	7	4,94	1,434
excur	18	2	7	4,61	1,685
quitatie	18	2	7	4,56	1,423
nomeint	18	1	4	2,61	1,145
Valid N (listwise)	18				

Grupo 2

	N	Minimum	Maximum	Mean	Std. Deviation
divertid	7	1	2	1,57	,535
pidocomp	7	1	3	1,86	,690
aprendom	7	2	4	3,29	,756
excur	7	2	5	3,43	,976
quitatie	7	2	4	2,86	,690
nomeint	7	4	7	5,29	1,113
Valid N (listwise)	7				

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

Grupo 1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4	22,2	22,2	22,2
	1	14	77,8	77,8	100,0
	Total	18	100,0	100,0	

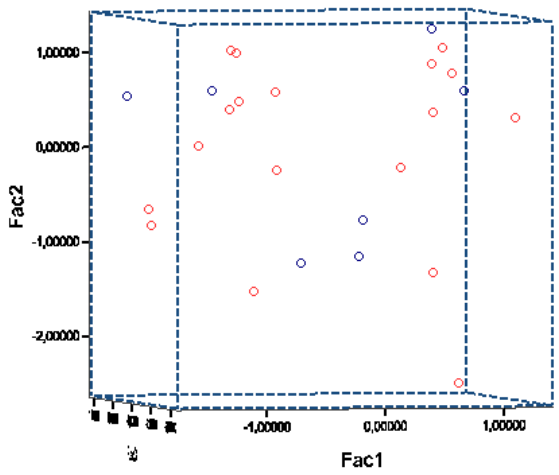
Grupo 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5	71,4	71,4	71,4
	1	2	28,6	28,6	100,0
	Total	7	100,0	100,0	

Componentes Principales

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,866	47,761	47,761	2,866	47,761	47,761
2	1,772	29,532	77,293	1,772	29,532	77,293
3	,863	14,384	91,677	,863	14,384	91,677
4	,382	6,374	98,052			
5	,095	1,590	99,642			
6	,021	,358	100,000			



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 (X_1), del área de metodología (X_2), del área de evolutiva (X_3), del área de social (X_4) y del área de clínica (X_5). Para ello hemos realizado la media por área y hemos obtenido la siguiente matriz:

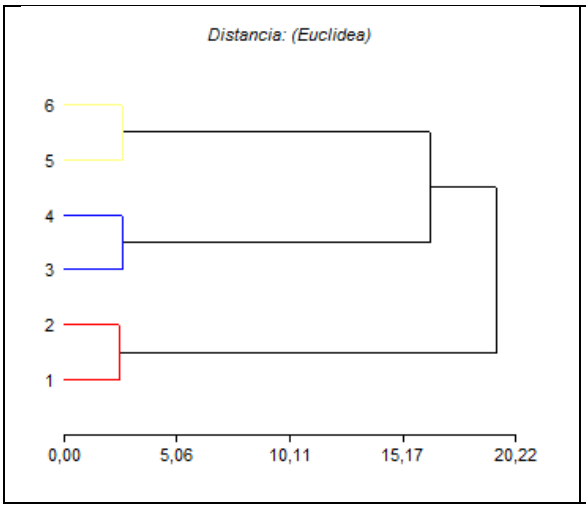
	X_1	X_2	X_3	X_4	X_5
S_1	8	9	7	8	6
S_2	7	8	7	8	8
S_3	2	3	8	7	2
S_4	1	2	6	7	1
S_5	1	1	1	9	8
S_6	2	3	1	8	9

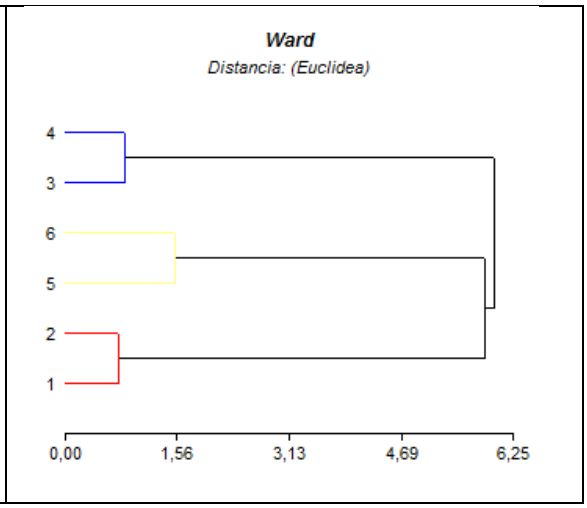
Con los datos de la matriz anterior realizar los dendogramas, utilizando el método de Ward:

a) Para los datos crudos
b) Para los datos estandarizados por variable
c) ¿A qué se deben las diferencias observadas en los dendrogramas?
d) 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

Euclidea

	1	2	3	4	5	6
1	0,00					
2	2,45	0,00				
3	19,38	9,38	0,00			
4	14,03	11,09	2,65	0,00		
5	19,25	11,05	16,42	11,44	0,00	
6	12,61	9,33	9,95	9,59	2,65	0,00

Euclidea

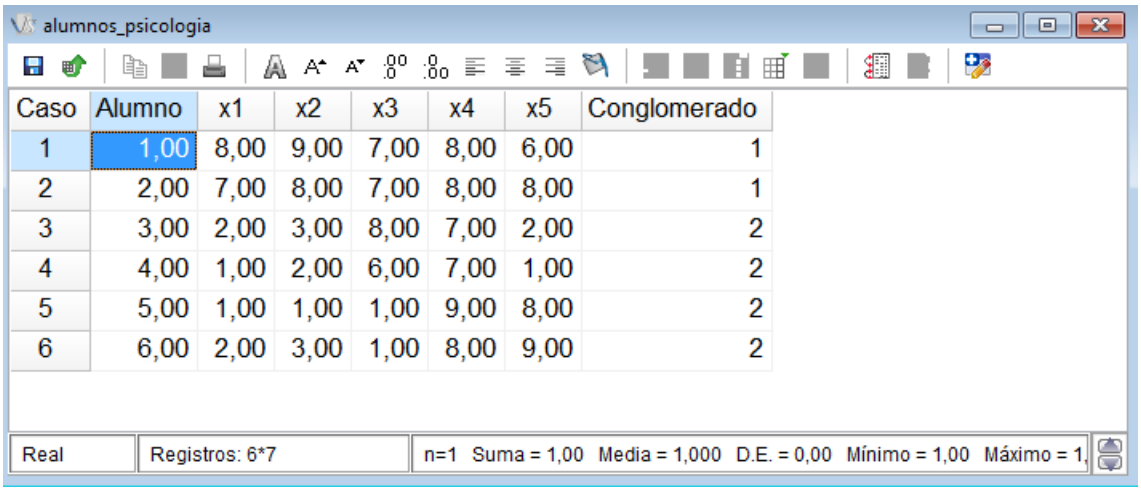
	1	2	3	4	5	6
1	0,00					
2	0,73	0,00				
3	5,99	3,13	0,00			
4	4,60	3,61	0,82	0,00		
5	5,87	3,66	5,90	3,73	0,00	
6	3,93	2,91	4,04	3,17	1,52	0,00

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.



Caso	Alumno	x1	x2	x3	x4	x5	Conglomerado
1	1,00	8,00	9,00	7,00	8,00	6,00	1
2	2,00	7,00	8,00	7,00	8,00	8,00	1
3	3,00	2,00	3,00	8,00	7,00	2,00	2
4	4,00	1,00	2,00	6,00	7,00	1,00	2
5	5,00	1,00	1,00	1,00	9,00	8,00	2
6	6,00	2,00	3,00	1,00	8,00	9,00	2

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

a- Utilizando el método de Ward y la distancia euclidea particionar en dos clusters. Como llamaría a cada uno de ellos?

b- Idem a- pero en cuatro clusters. Utilizando el dendograma, con cuál de las clasificaciones se quedaría?

c- Realice una cauterización de las variables?.

d- Compare los resultados obtenidos con el de componentes principales.