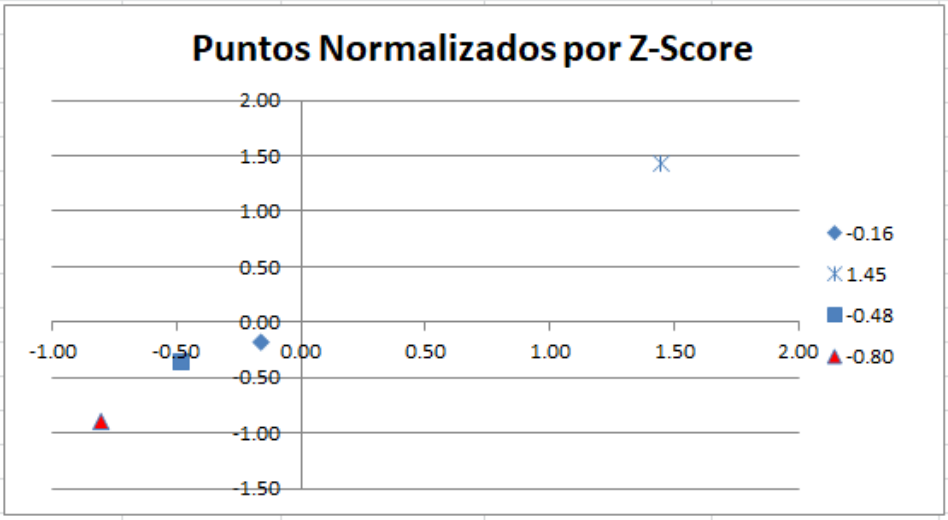
TRABAJO PRÁCTICO V: Minería de datos

PARTE 02: Clustering (K-Medias y algoritmos jerárquicos)

1. Medidas de distancia. Calcule la distancia entre los siguientes puntos y el centroide (2, 4) utilizando las medidas: euclídea, Manhattan y Minkowski (con p = 3):

		normalizado por Z-Score (X-media(X))/sd(X)		distancias respecto del punto (2,4)		
X	Y	X	Y	distancia euclidea	distancia Manhattan	distancia Minkowski
4	8	-0.16	-0.18	0.96	1.36	0.79
9	17	1.45	1.43	3.23	4.57	4.89
3	7	-0.48	-0.36	0.63	0.86	0.43
2	4	-0.80	-0.89	0.00	0.00	0.00

mean X	4.50
mean Y	9.00
sd X	3.11
sd Y	5.60



Se puede observar que, dado el centroide: (2,4) el punto mas cercano es el punto: (3,7), esto se ve con el calculo de distancias, para el cual se obtienen los valores:

- 0.63 (euclideana)
- 0.86 (Manhattan)
- 0.43 (Minkowski)

Que son los valores mas chicos, coincidiendo con la grafica.

2. A continuación, calcule la distancia entre las diferentes variables de tipo categóricas con respecto a la instancia {1, lluvioso, templado, alta, fuerte}:

basandonos en la formula para el calculo de distancias con variables categoricas, armamos la matriz binaria correspondiente.

$$d(i, j) = \frac{p - m}{p},$$

1	lluvioso	templado	alta	fuerte	
					DISTANCIA RESPECTO DE #1
2	lluvioso	frio	normal	fuerte	0.5
3	nublado	frio	normal	fuerte	0.75
4	soleado	templado	alta	leve	0.5
5	soleado	frio	normal	leve	1
6	lluvioso	templado	normal	leve	0.5
7	soleado	templado	normal	fuerte	0.5
8	nublado	templado	alta	fuerte	0.25
9	nublado	calor	normal	leve	1
10	lluvioso	templado	alta	fuerte	0

Se puede observar que los valores mas cercanos a cero, seran los mas cercanos a la instancia #1, en este caso las instancias #8 y #10.

3. Ahora, y a partir de los datos de la siguiente tabla, agrupe los datos de acuerdo al algoritmo k-medias utilizando la medida euclídea y con los puntos A1, A2 y A7 como centroides iniciales.

Resolucion:

a). normalizo por z-score

				normalizado por Z-Score (X-media(X))/sd(X)	
PUNTOS	X	Y		X	Y
A1	2	10		-0.75	1.57
A2	2	5		-0.75	-0.28
A3	8	4		1.50	-0.65
A4	2	7		-0.75	0.46
A5	7	5		1.12	-0.28
A6	6	4		0.75	-0.65
A7	1	2		-1.12	-1.38
A8	4	9		0.00	1.20

b). establezco centroides iniciales

	#	CENTROIDES		
ITERACION #0	K1	-0.75	1.57	centroides iniciales
	K2	-0.75	-0.28	
	K3	-1.12	-1.38	

c). calculo iteracion 0

normalizado por Z-Score (X-media(X))/sd(X)			ITERACION #0		
X	Y	ITERACION #0	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3
-0.75	1.57	K1	0.00	1.84	2.97
-0.75	-0.28	K2	1.84	0.00	1.17
1.50	-0.65	K2	3.15	2.28	2.72
-0.75	0.46	K2	1.11	0.74	1.88
1.12	-0.28	K2	2.63	1.87	2.50
0.75	-0.65	K2	2.67	1.54	2.01
-1.12	-1.38	K3	2.97	1.17	0.00
0.00	1.20	K1	0.83	1.65	2.81

c-1). actualizo centroides

	#	CENTROIDES		
ITERACION #0	K1	-0.75	1.57	centroides iniciales
	K2	-0.75	-0.28	
	K3	-1.12	-1.38	
ITERACION #1	K1	-0.37	1.38	
	K2	0.37	-0.28	
	K3	-1.12	-1.38	

d). calculo iteracion 1

normalizado por Z-Score (X-media(X))/sd(X)				ITERACION #0			ITERACION #1		
X	Y	ITERACION #0	ITERACION #1	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3
-0.75	1.57	K1	K1	0.00	1.84	2.97	0.42	2.16	2.97
-0.75	-0.28	K2	K2	1.84	0.00	1.17	1.70	1.12	1.17
1.50	-0.65	K2	K2	3.15	2.28	2.72	2.76	1.18	2.72
-0.75	0.46	K2	K1	1.11	0.74	1.88	0.99	1.34	1.88
1.12	-0.28	K2	K2	2.63	1.87	2.50	2.23	0.75	2.50
0.75	-0.65	K2	K2	2.67	1.54	2.01	2.32	0.53	2.01
-1.12	-1.38	K3	K3	2.97	1.17	0.00	2.86	1.86	0.00
0.00	1.20	K1	K1	0.83	1.65	2.81	0.42	1.52	2.81

d-1). actualizo centroides

	#	CENTROIDES		
ITERACION #0	K1	-0.75	1.57	centroides iniciales
	K2	-0.75	-0.28	
	K3	-1.12	-1.38	
ITERACION #1	K1	-0.37	1.38	
	K2	0.37	-0.28	
	K3	-1.12	-1.38	
ITERACION #2	K1	-0.50	1.08	
	K2	0.65	-0.46	
	K3	-1.12	-1.38	

e). calculo iteracion 2

normalizado por Z-Score (X-media(X))/sd(X)					ITERACION #0			ITERACION #1			ITERACION #2		
X	Y	ITERACION #0	ITERACION #1	ITERACION #2	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3
-0.75	1.57	K1	K1	K1	0.00	1.84	2.97	0.42	2.16	2.97	0.55	2.47	2.97
-0.75	-0.28	K2	K2	K3	1.84	0.00	1.17	1.70	1.12	1.17	1.37	1.42	1.17
1.50	-0.65	K2	K2	K2	3.15	2.28	2.72	2.76	1.18	2.72	2.63	0.86	2.72
-0.75	0.46	K2	K1	K1	1.11	0.74	1.88	0.99	1.34	1.88	0.66	1.68	1.88
1.12	-0.28	K2	K2	K2	2.63	1.87	2.50	2.23	0.75	2.50	2.11	0.50	2.50
0.75	-0.65	K2	K2	K2	2.67	1.54	2.01	2.32	0.53	2.01	2.13	0.21	2.01
-1.12	-1.38	K3	K3	K3	2.97	1.17	0.00	2.86	1.86	0.00	2.54	2.00	0.00
0.00	1.20	K1	K1	K1	0.83	1.65	2.81	0.42	1.52	2.81	0.51	1.78	2.81

e-1). actualizo centroides

	#	CENTROIDES		
ITERACION #0	K1	-0.75	1.57	centroides iniciales
	K2	-0.75	-0.28	
	K3	-1.12	-1.38	
ITERACION #1	K1	-0.37	1.38	
	K2	0.37	-0.28	
	K3	-1.12	-1.38	
ITERACION #2	K1	-0.50	1.08	
	K2	0.65	-0.46	
	K3	-1.12	-1.38	
ITERACION #3	K1	-0.50	1.08	
	K2	1.12	-0.52	
	K3	-0.94	-0.83	

f). calculo iteracion 3

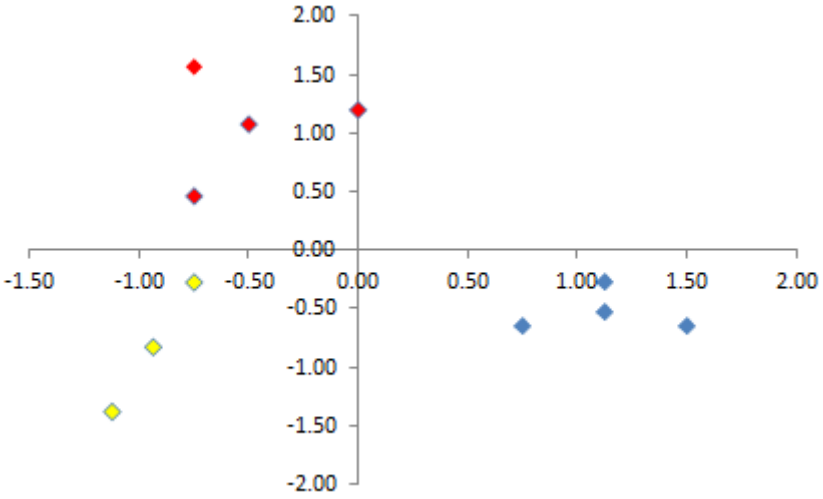
normalizado por Z-Score (X-media(X))/sd(X)						ITERACION #0			ITERACION #1			ITERACION #2			ITERACION #3		
X	Y	ITERACION #0	ITERACION #1	ITERACION #2	ITERACION #3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3	distancia euclidea RESPECTO DE #K1	distancia euclidea RESPECTO DE #K2	distancia euclidea RESPECTO DE #K3
-0.75	1.57	K1	K1	K1	K1	0.00	1.84	2.97	0.42	2.16	2.97	0.55	2.47	2.97	0.55	2.80	2.40
-0.75	-0.28	K2	K2	K3	K3	1.84	0.00	1.17	1.70	1.12	1.17	1.37	1.42	1.17	1.37	1.89	0.58
1.50	-0.65	K2	K2	K2	K2	3.15	2.28	2.72	2.76	1.18	2.72	2.63	0.86	2.72	2.63	0.39	2.44
-0.75	0.46	K2	K1	K1	K1	1.11	0.74	1.88	0.99	1.34	1.88	0.66	1.68	1.88	0.66	2.11	1.30
1.12	-0.28	K2	K2	K2	K2	2.63	1.87	2.50	2.23	0.75	2.50	2.11	0.50	2.50	2.11	0.25	2.13
0.75	-0.65	K2	K2	K2	K2	2.67	1.54	2.01	2.32	0.53	2.01	2.13	0.21	2.01	2.13	0.39	1.69
-1.12	-1.38	K3	K3	K3	K3	2.97	1.17	0.00	2.86	1.86	0.00	2.54	2.00	0.00	2.54	2.40	0.58
0.00	1.20	K1	K1	K1	K1	0.83	1.65	2.81	0.42	1.52	2.81	0.51	1.78	2.81	0.51	2.05	2.23

ya no hay cambios en los clusters, por lo tanto finalizo el algoritmo.

Centroides Finales y grafica de puntos

X	Y
-0.75	1.57
-0.75	-0.28
1.50	-0.65
-0.75	0.46
1.12	-0.28
0.75	-0.65
-1.12	-1.38
0.00	1.20

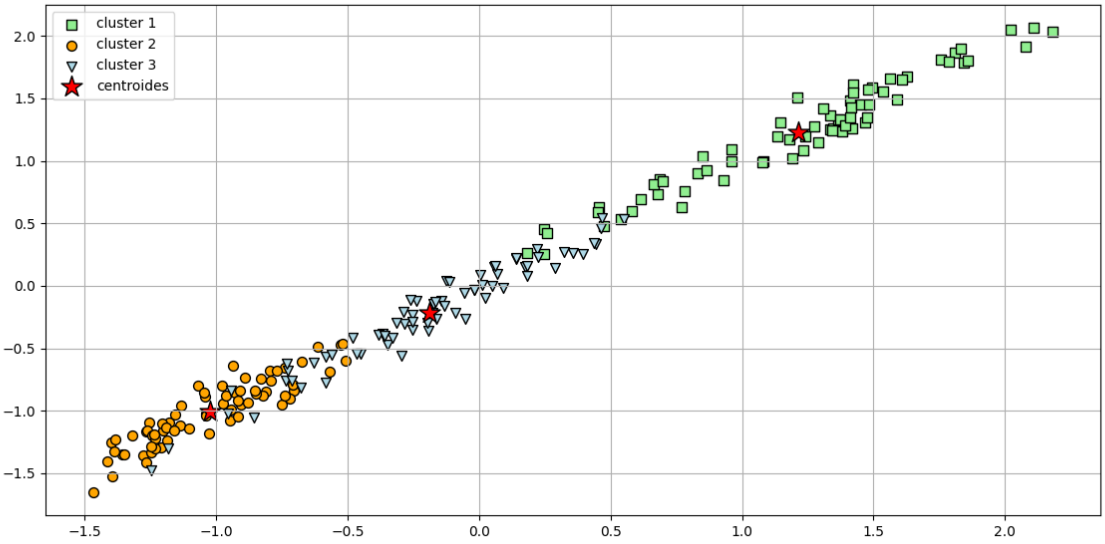
CENTROIDES	
-0.50	1.08
1.12	-0.52
-0.94	-0.83



Conclusion: Se pueden observar tres cluster bien definidos.

4. K-means. Se provee un dataset sobre las características internas del núcleo de tres clases de trigo diferentes. Cargue el dataset en una de las herramientas de minería de datos provistas y resuelva:

- a) Utilice el algoritmo k-medias variando la cantidad de centroides a efectos de agrupar los datos de la manera más eficiente.



- b) ¿Cuál es la cantidad de grupos que permite un mejor agrupamiento de los datos? ¿Mediante cual métrica puede verificar esto?

La metrica que me permite verificar el mejor agrupamiento de los datos es el coeficiente de silueta

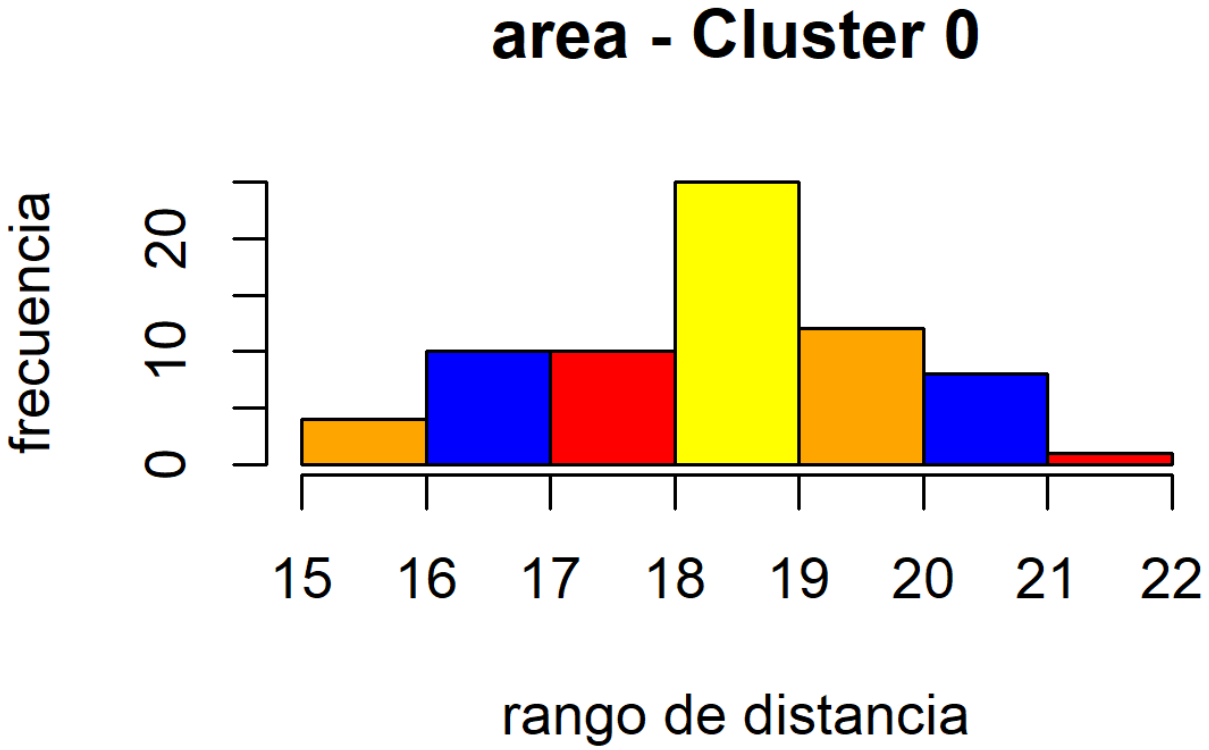
```
For n_clusters =2 silhouette score is 0.4039111372552087
For n_clusters =3 silhouette score is 0.4850201995176518
For n_clusters =4 silhouette score is 0.38957966159931
For n_clusters =5 silhouette score is 0.2892214039235866
For n_clusters =6 silhouette score is 0.25909420201695776
```

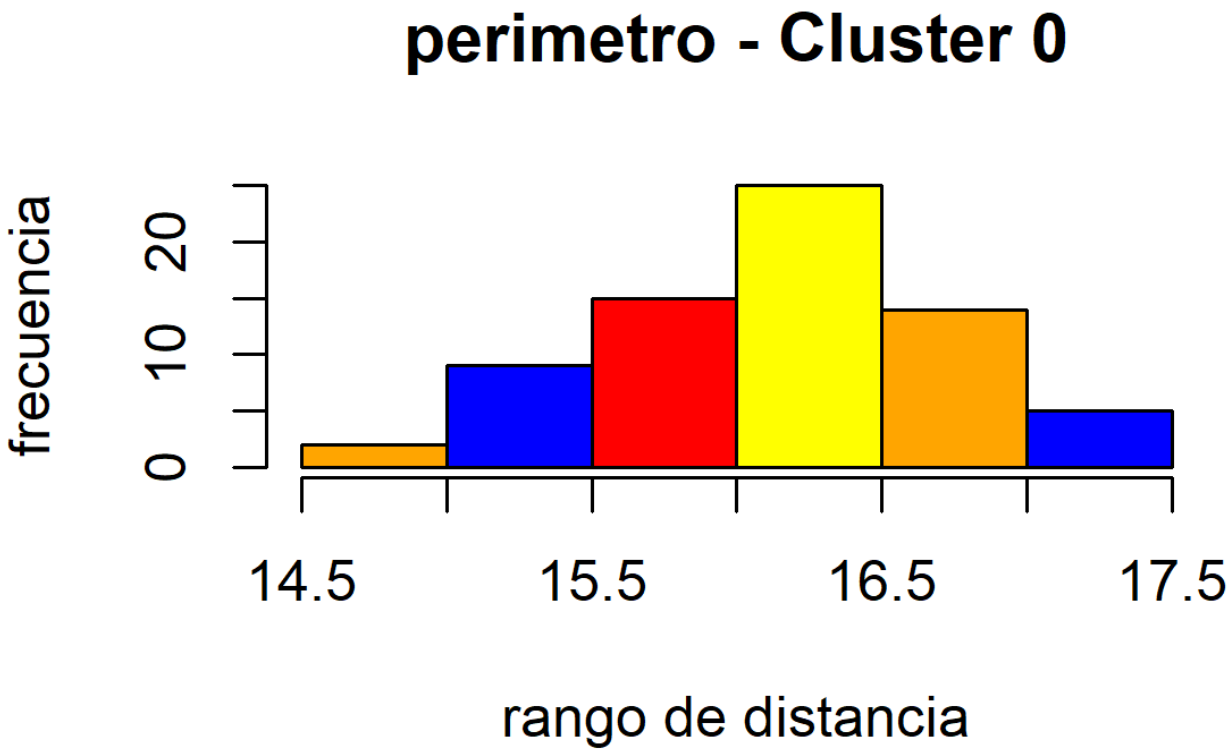
Se puede observar que un K=3 es la cantidad optima de agrupamientos.

- c) ¿Cuáles son las características más distintivas de cada uno de los cluters resultantes?

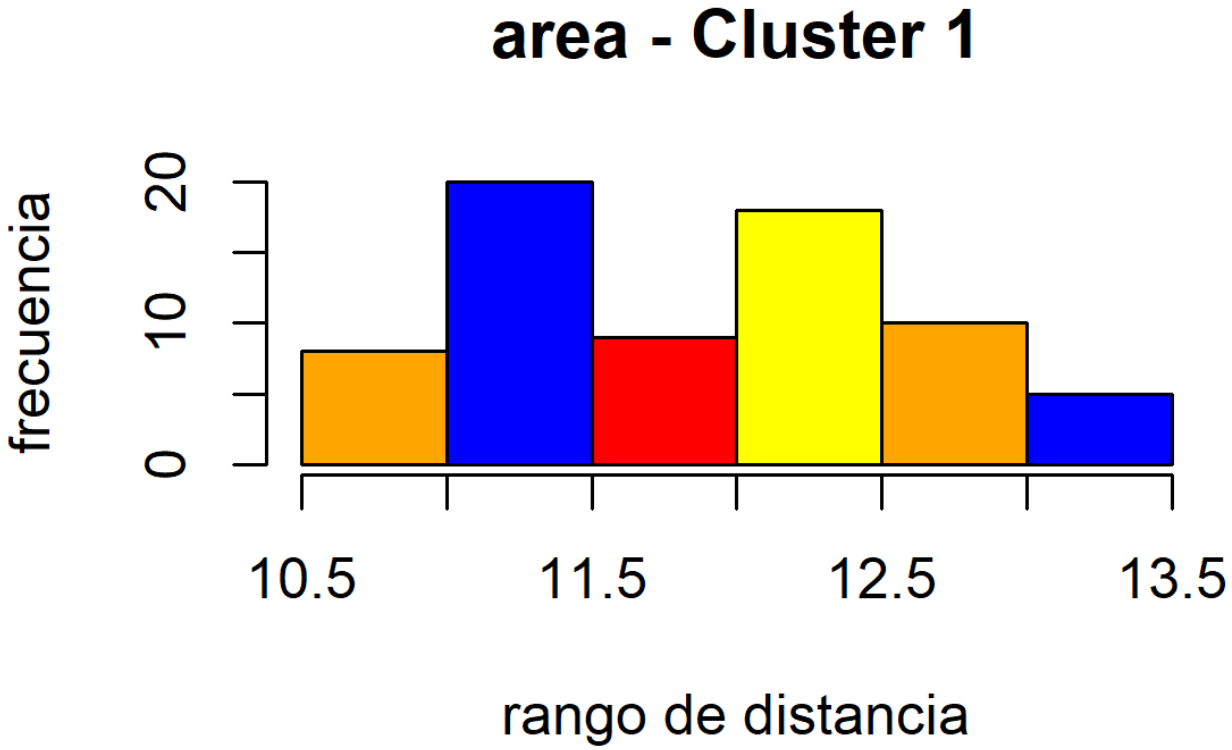
Realizando un analisis exploratorio de cada cluster,

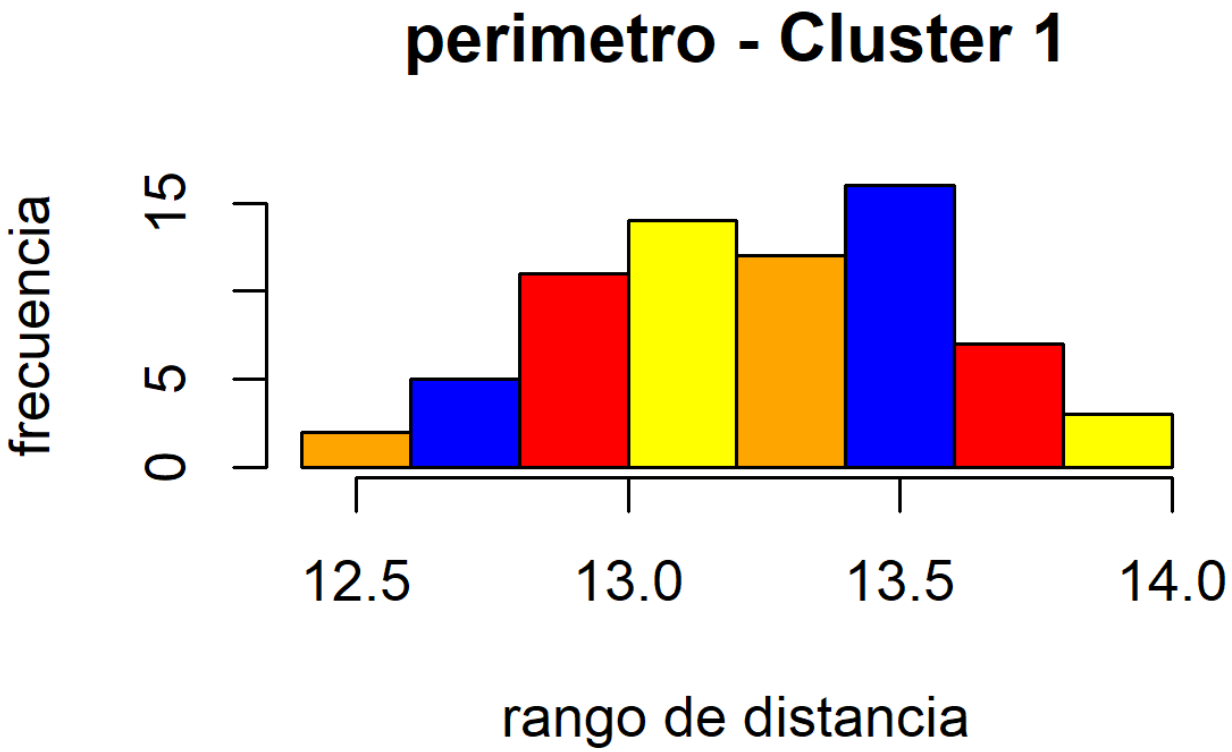
- Las caracteristicas distintivas son "el area", "el perimetro", "compact", "la longitud del kernel" y "el ancho del kernel", de estos solo analice los principales: area y perimetro. En base a los graficos pude observar lo siguiente:



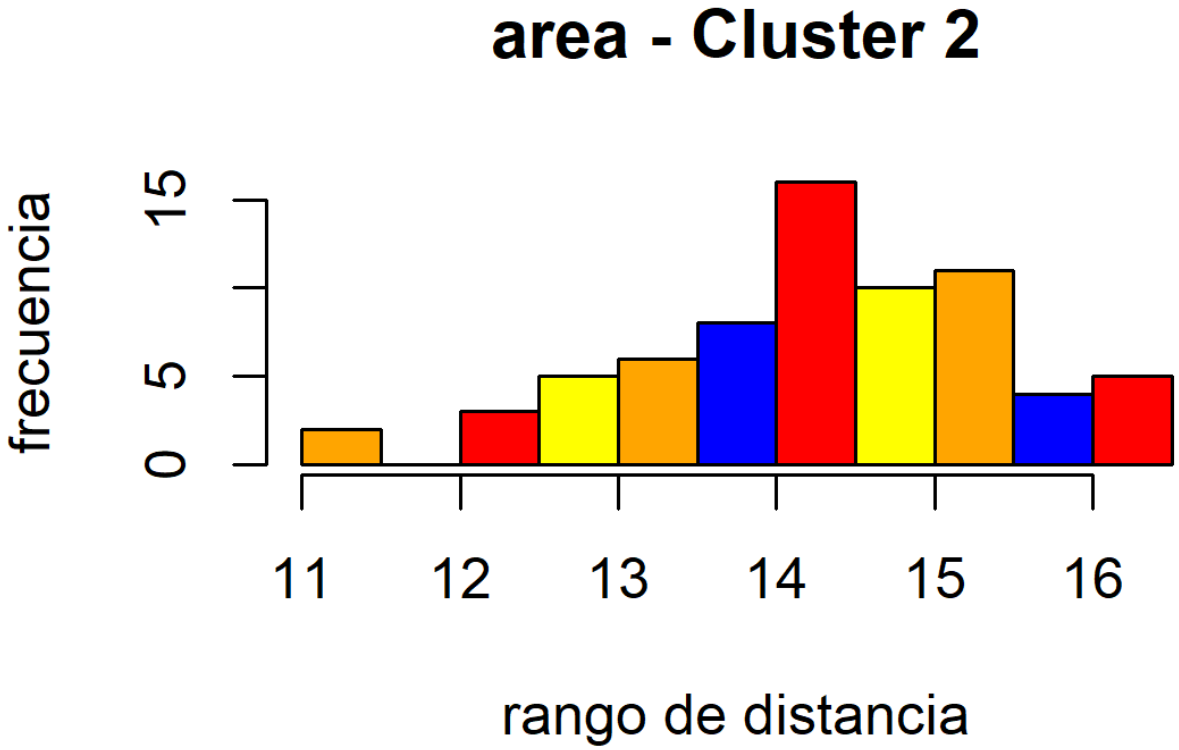


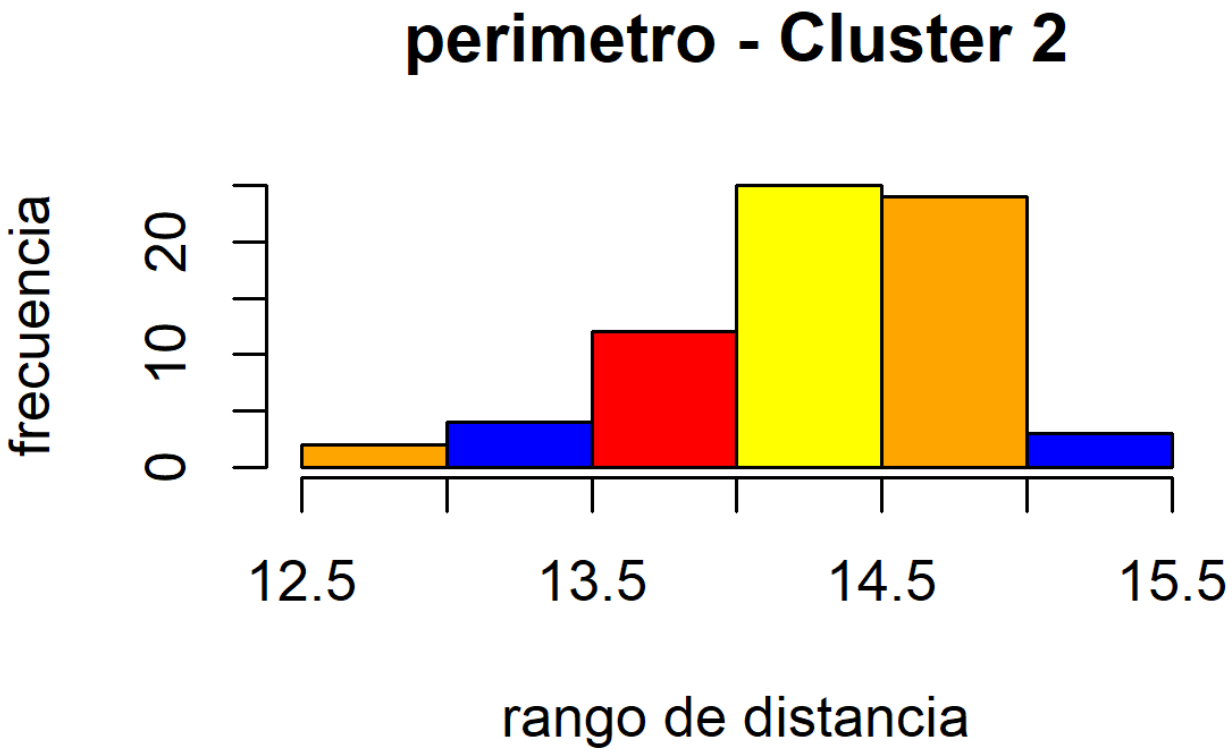
- Se puede observar un area entre 18 ~ 19
- Y un perimetro entre 16 ~ 16,5





- Se puede observar un area entre 11 ~ 11,5
- Y un perimetro entre 13 ~ 13,5





- Se puede observar un area entre 14 ~ 14,5
- Y un perimetro entre 14 ~ 15

5. Ahora, trabaje sobre el dataset abandono_cuantitativo.csv:

- a) Escoja los features que a su entender permitan un mejor agrupamiento, pre-procese los mismos y entrene un modelo a partir de K-Means.

En base a un Analisis de Componentes Principales, puedo observar que las variables que describen mejor el dataset son:

```
    'c_noausentes_1er_anio',
    'cursadas_ap_1er_anio',
    'c_libres_1er_anio',
    'cursadas_1er_anio',
    'fracaso_academico',
```

```
                              PC1          PC2          PC3          PC4
X1                      0.03654832  0.040168817  0.001148992  0.11936271
horas_trabajadas        0.11792527 -0.245122404  0.236029534 -0.46406028
edad_ingreso            0.16365758 -0.294331286  0.320067466 -0.40786687
colegio_publico         0.12929108 -0.027202106  0.169602560 -0.51109782 -
aprobadas_1er_anio     -0.39919786 -0.201495446  0.119776498  0.06321085
c_promociones_1er_anio -0.35944268 -0.238638702  0.119122040  0.09679169
c_libres_1er_anio      -0.14054523  0.608908275 -0.028427085 -0.28122075 -
c_regulares_1er_anio   -0.32682379 -0.002955144  0.005428794 -0.16266734
c_ausentes_1er_anio     0.13958011 -0.096139201 -0.734156610 -0.29328466
c_noausentes_1er_anio  -0.42791138  0.127260418  0.043064162 -0.15826018
cursadas_ap_1er_anio   -0.42763589 -0.163520432  0.083848992 -0.02585250
cursadas_1er_anio      -0.36946922  0.080621605 -0.348498796 -0.32137374
cambio_universidad     -0.04918023 -0.031085153 -0.002065862  0.07839098 -
fracaso_academico       0.04175804  0.568296336  0.338200688 -0.05251154 -
```

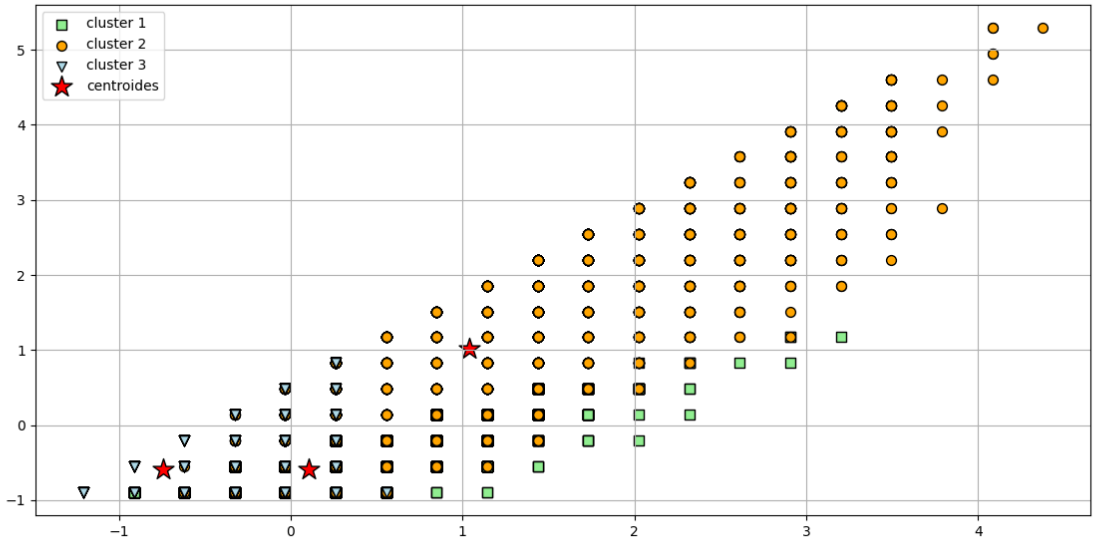
```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.2477 1.3479 1.13585 1.08292 1.02346 0.97038 0.93141 0.87120 0.78049
Proportion of Variance 0.3609 0.1298 0.09215 0.08377 0.07482 0.06726 0.06197 0.05421 0.04351
Cumulative Proportion  0.3609 0.4906 0.58281 0.66657 0.74139 0.80865 0.87062 0.92483 0.96834
                          PC10   PC11    PC12      PC13      PC14
Standard deviation     0.60580 0.2510 0.11497 3.219e-15 1.418e-15
Proportion of Variance 0.02621 0.0045 0.00094 0.000e+00 0.000e+00
Cumulative Proportion  0.99456 0.9991 1.00000 1.000e+00 1.000e+00
```

- Los features escogidos son

```
    'c_noausentes_1er_anio',
    'cursadas_ap_1er_anio',
    'c_libres_1er_anio',
    'cursadas_1er_anio',
    'fracaso_academico',
```

A partir del cual, se obtiene un modelo que tiene cluster superpuestos. Podria seguir acotando el analisis, eliminando mas variables y subiria el coeficiente de silueta al 60%, pero me parecio que estaba acotando demasiado y estaria sacando del dataset variables que explican una parte importante del mismo, en pos de

buscar un mejor agrupamiento de los datos. La idea tampoco es perder imformacion, asi que elegi quedarme con el 50% de coeficiente de silueta y 3 grupos.



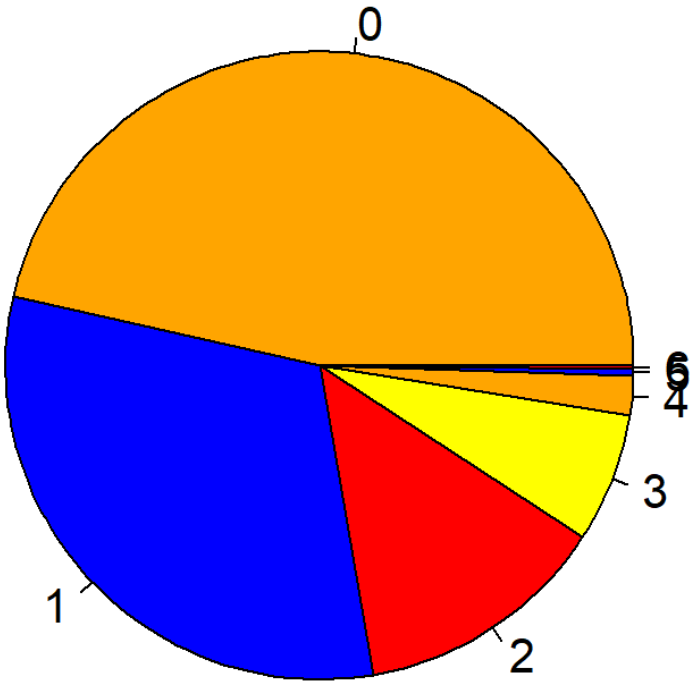
- b) Analice y describa las características más salientes de cada uno de los grupos encontrados por el algoritmo.

De los grupos resultantes analice dos variables principales:

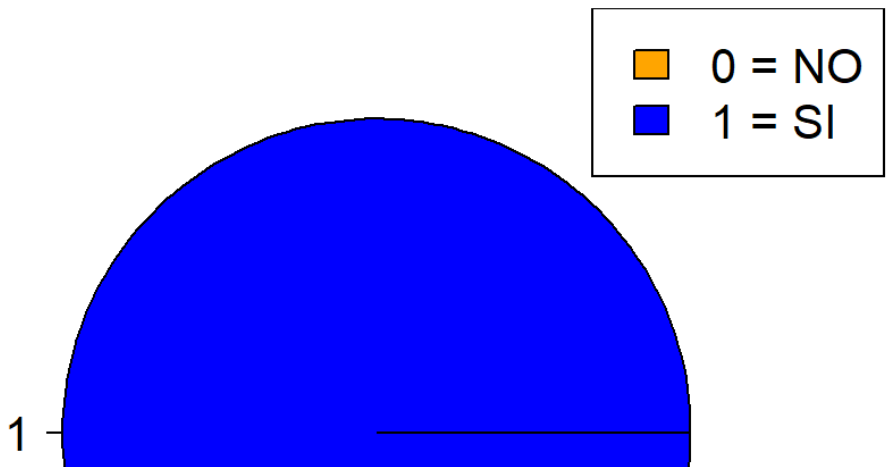
- 'cursadas_ap_1er_anio' y
- 'fracaso_academico',

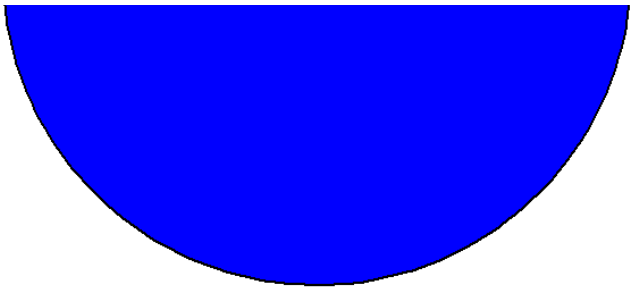
Y a partir de un analisis grafico de los datos, realice las siguientes observaciones.

cursadas_ap_1er_anio - Cluster 0



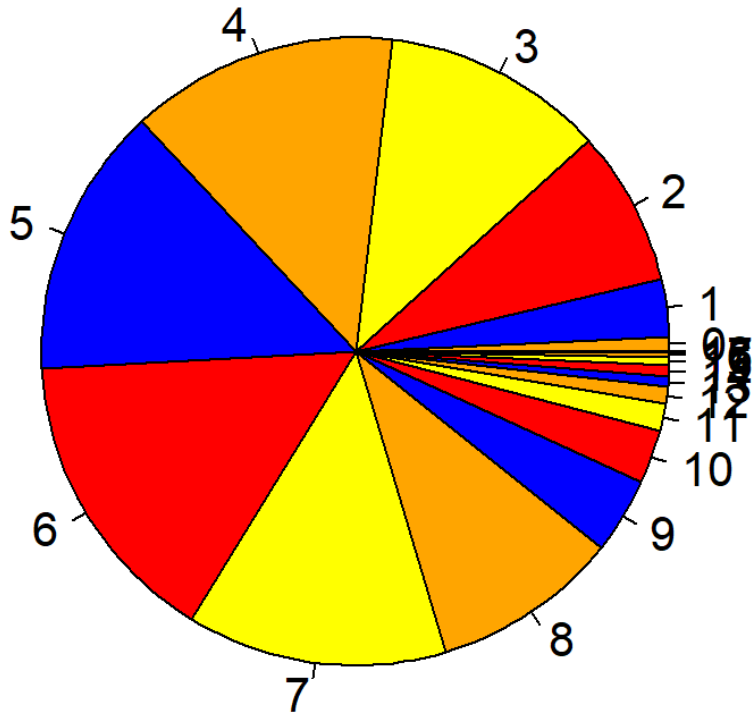
Fracaso Academico - Cluster 0



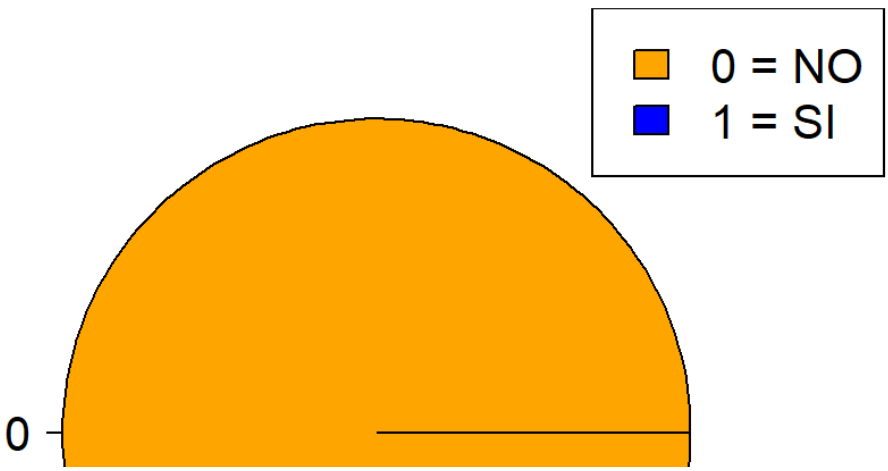


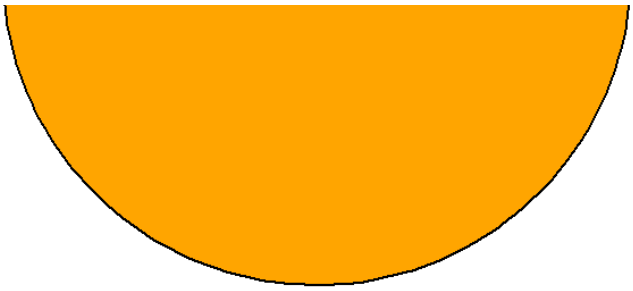
- Se puede observar que: se encuentran aquellos estudiantes que mayormente, no cursaron materias el primer 1er año y fracasaron academicamente.

cursadas_ap_1er_anio - Cluster 1



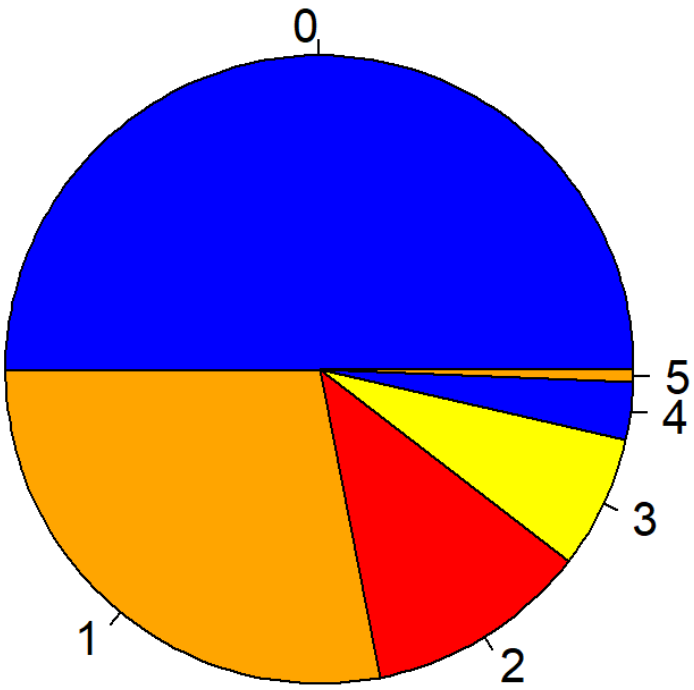
Fracaso Academico - Cluster 1



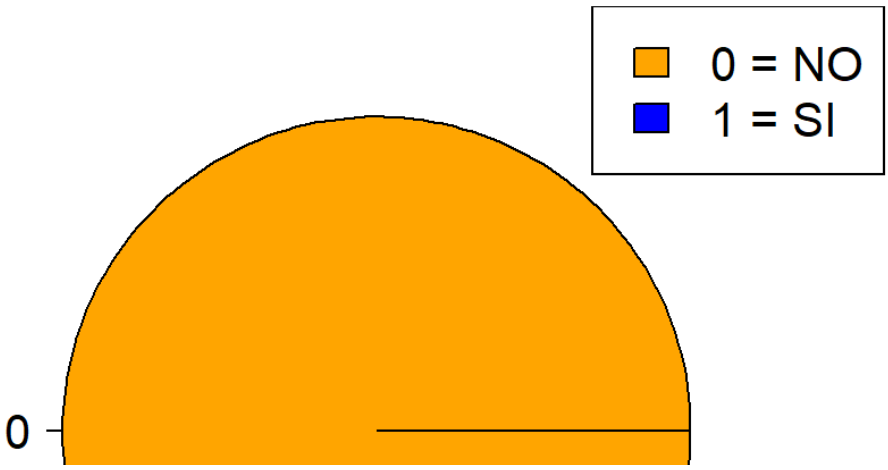


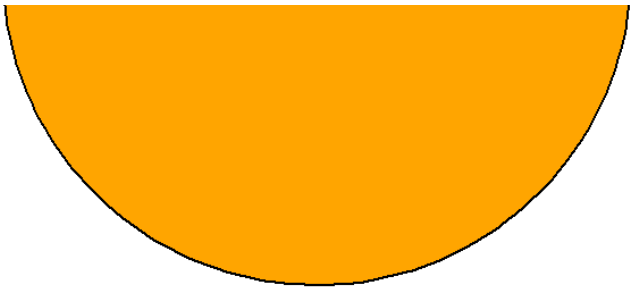
- El segundo grupo: hay estudiantes que cursaron en promedio 2 materias y ninguno fracaso academicamente.

cursadas_ap_1er_anio - Cluster 2



Fracaso Academico - Cluster 2





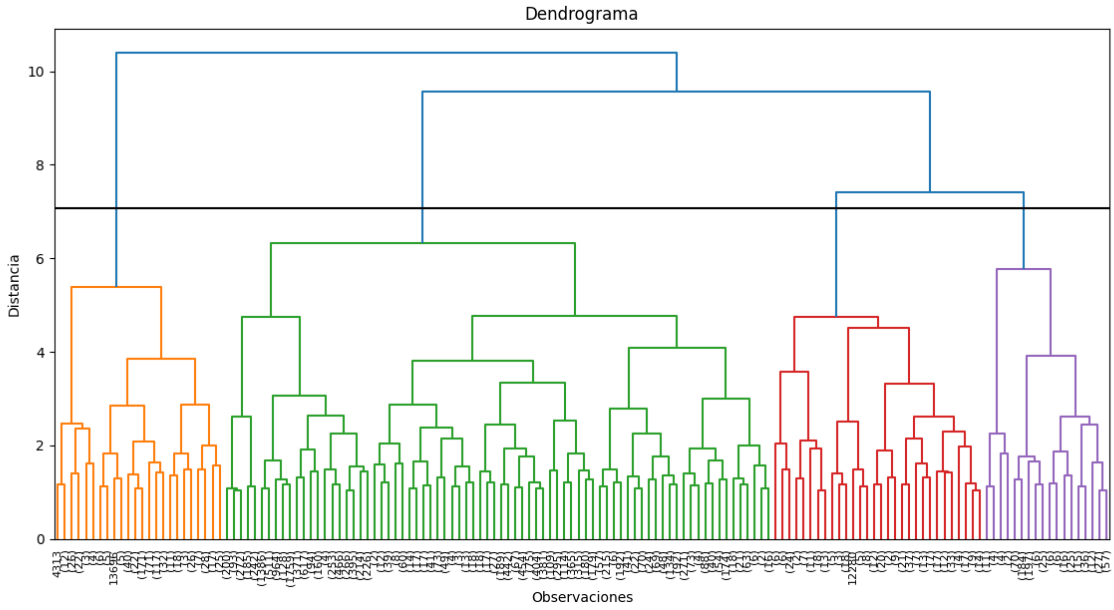
- Por ultimo, este cluster es parecido al cluster 0, con la diferencia que tenemos estudiantes que no fracasaron academicamente a pesar de que, en su mayoria no cursaron materias el primer año.

- c) Encuentre la cantidad de grupos que logran el mejor agrupamiento para los datos. Justifique la elección a partir de métricas y gráficas de los conglomerados resultantes.

cant_cluster	coeficiente_silueta
2	0.4172791289245163
3	0.5039014903639526
4	0.43057631276938785
5	0.44976901761026955

Segun la metrica de silueta se puede observar que tres cluster logran el mejor agrupamiento.

- d) Ahora aplique algún algoritmo jerárquico a efectos de agrupar los datos. ¿Cuál nivel se corresponde con el agrupamiento realizado por k-medias en el punto 6) a)?



Se puede observar que, a una distancia aproximada de 7 corresponde una agrupancion de k=4. A una distancia aproximada de 8 se corresponde con el agrupamiento realizado por k-medias en el punto 6) a).

- e) ¿El agrupamiento jerárquico permite encontrar una mejor forma de agrupar los datos? Si fuera así, ¿Cuál es ese agrupamiento?

Dependiente del tipo de agrupamiento utilizado, se puede obtener un mejor agrupamiento, en el caso anterior se uso un metodo de linkage de `Complete Linkage`, lo que permitio observar un agrupamiento mas uniforme.

6. Algoritmos jerárquicos. Incorpore en Colab nuevamente el dataset del punto 5 y realice las siguientes actividades:

- a) Realice el agrupamiento de los datos utilizando diferentes parámetros.

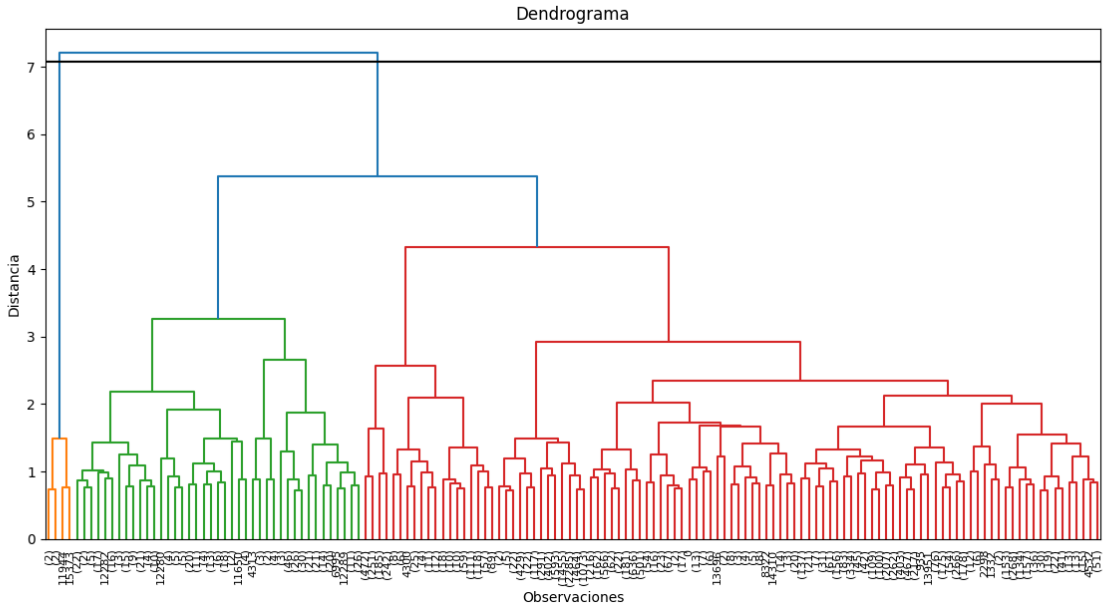
Variando el parametro del tipo de linkage se pueden realizar dintintos agrupamientos: Como ser `single`, `average`, `centroid` y `complete`

```
from scipy.cluster.hierarchy import dendrogram, linkage

# H = linkage(scaled_1, 'single')
# H = linkage(scaled_1, 'average')
# H = linkage(scaled_1, 'centroid')
H = linkage(scaled_1, 'complete')

from scipy.spatial.distance import pdist, squareform

squareform(pdist(scaled_1))

max_d = 7.08
plt.figure(figsize=(25, 10))
plt.title('Dendrograma')
plt.xlabel('Observaciones')
```

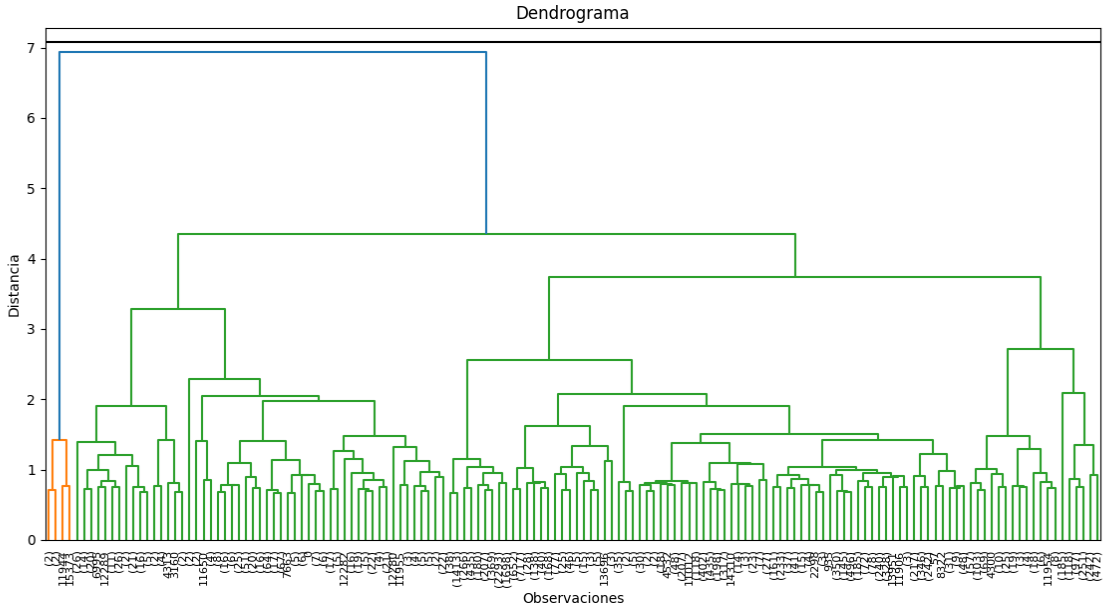
```
plt.ylabel('Distancia')
dendrogram(
    H, truncate_mode='mlab',
    p=5, leaf_rotation=90.,
    leaf_font_size=8.,
)
plt.axhline(y=max_d, c='k')
plt.show()
```

- b) Grafique el resultado y escoja cual es el nivel que mejor agrupa los datos.

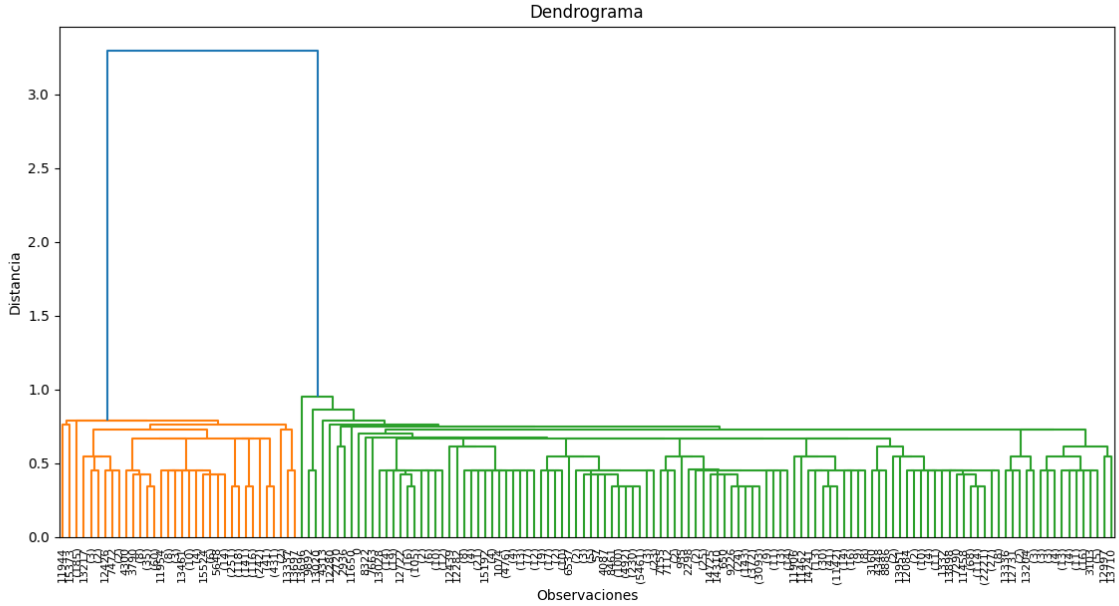
Dendograma AVERAGE



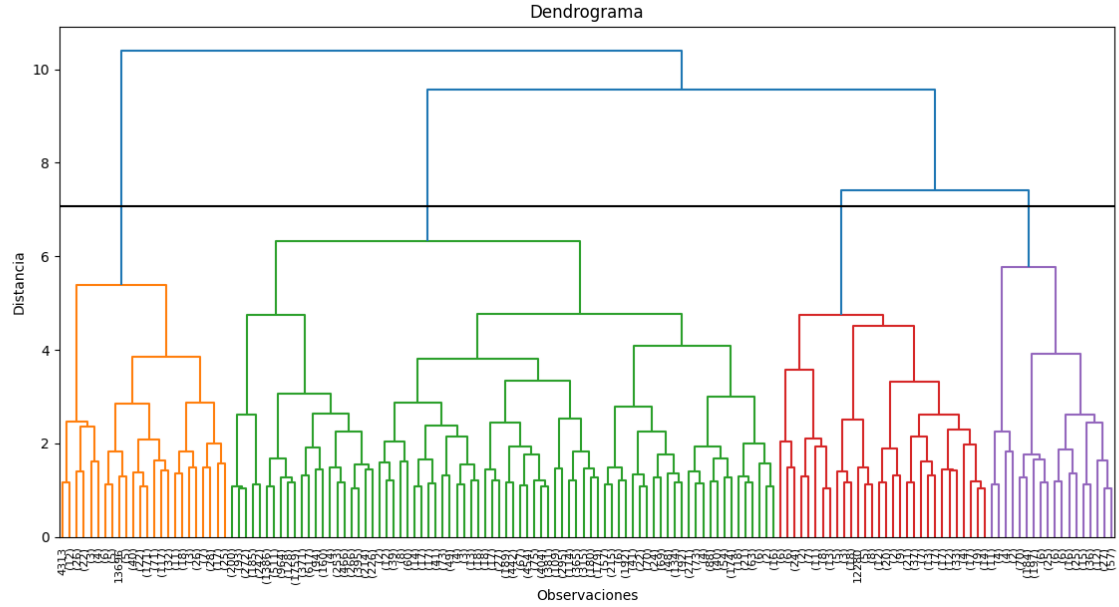
Dendograma CENTROID



Dendograma `SINGLE`



Dendograma `COMPLETE`



Se puede observar que el dendograma `Complete` es el que mejor agrupa los datos.