Inteligencia de Negocio (2019-2020)
Grado en Ingeniería Informática
Universidad de Granada

Memoria Práctica 2
Análisis Relacional mediante Segmentación

Daniel Terol Guerrero
DNI: 09076204J
Correo: danielterol@correo.ugr.es

6 de diciembre de 2019

Índice

1. Introducción

En esta segunda práctica se ha abordado un problema de segmentación mediante el uso de técnicas de aprendizaje no supervisado. Los métodos utilizados son los siguientes:

- Varios **métodos de particionamiento** como son, por ejemplo, KMeans, MeanShift, DBSCAN y Birch.
- Un método aglomerativo que sigue una estrategia **Ward**

El conjunto de datos sobre el que se ha realizado la segmentación consiste en las respuestas de diferentes mujeres a un test sobre fecundidad. En total, se dispone de 14.556 respuestas con 463 variables sobre datos personales, datos biográficos, hogar, padres, relaciones de pareja, estudios, empleo, hijos, creencias, etc.

El estudio que se va a realizar consta de tres casos de estudio:

1. Satisfacción de las mujeres trabajadoras y embarazadas recientemente en el reparto de tareas con su pareja.
2. Estudios alcanzados y campos en los que se han especializado las mujeres.
3. Resentimiento de las mujeres por no haber tenido hijos/no haber tenido tantos como querían.

Cada caso de estudio será segmentado por cada técnica mostrando las diferentes agrupaciones del algoritmo en gráficas como *HeatMaps, ScatterMatrix o Dendograma*. Además, los métodos de particionamiento cuentan con medidas de rendimiento, como el *Coeficiente de Silhouette* y *Calinski-Harabasz*.

Se remarca que cada método tiene su propia configuración. Por ejemplo, KMeans requiere la especificación del número de clusters mientras que, por ejemplo, Meanshift estima, de forma autónoma, el número de clusters óptimo. Por tanto, al tener cada método su propia configuración, cada caso de estudio se realizará con una configuración diferente.

Por último, para poder estimar el número óptimo de clusters usando *KMeans*, se ha utilizado un método llamado *Elbow Method*.

Este método, ejecuta KMeans en el conjunto de datos para un rango de valores de clusters y, para cada valor, calcula una puntuación media. Dicha puntuación puede ser dos:

1. Distorsión: Se calcula como la media de las distancias al cuadrado desde los centroides de los diferentes clusters. Se suele usar la distancia Euclídea.

2. Inercia: Es la suma de las distancias al cuadrado de los microdatos a su centroide más cercano.

Aunque se mostrará la gráfica más adelante, el número óptimo de clusters es en el que, a partir de él, la gráfica decrece de forma lineal.

2. Caso de estudio 1: Satisfacción de las mujeres en el reparto de tareas con la pareja

```
subset = censo.loc[((censo['EDAD']>25) & (censo['EDAD']<=40))]
subset = subset.loc[(censo['EC']==2) & (censo['EMBANT']==1) & (censo['TRABAJAACT']==1)]

#seleccionar variables de interés para clustering
usadas = ['ESTUDIOSA', 'ESTUDIOSPAR', 'SITLABPAR', 'SATISFACENINOS']
```

(a) Código asociado al primer caso de estudio

Antes de comenzar a tratar el caso de estudio con las diferentes técnicas, se muestra el código correspondiente al caso de estudio y una explicación de por qué se han elegido esas variables. Las variables escogidas para establecer el subconjunto de datos es la siguiente:

- A través de la edad, se eligen las mujeres entre 25 y 40 años.
- La variable *EC* hace referencia al estado civil. Como me interesa las mujeres que estén casadas y, por tanto, tengan pareja actualmente, igualo esa variable al código 2, correspondiente con *Casado/a*.
- La variable *EMBANT* hace referencia a si la mujer ha estado embarazada recientemente. Por tanto, se iguala al código 1, correspondiente a la respuesta afirmativa.
- *TRABAJAACT* hace referencia si tiene un empleo actualmente. Al igual que *EMBANT*, *TRABAJAACT* se iguala al código 1, correspondiente con la respuesta afirmativa.

Por otro lado, las variables elegidas para realizar la segmentación son las siguientes:

- Tanto *ESTUDIOSA* y *ESTUDIOSPAR* son variables elegidas para ver si hay algún tipo de relación entre la satisfacción y el nivel de estudios alcanzados por ambas personas. Esta variable va desde *Menos que primaria* (código 1) y *Enseñanzas de doctorado* (código 9).
- *SITLABPAR* es una variable que representa la situación laboral de la pareja. Se ha elegido para ver, por ejemplo, si la pareja está parada y el grado de satisfacción es mayor (al no trabajar,

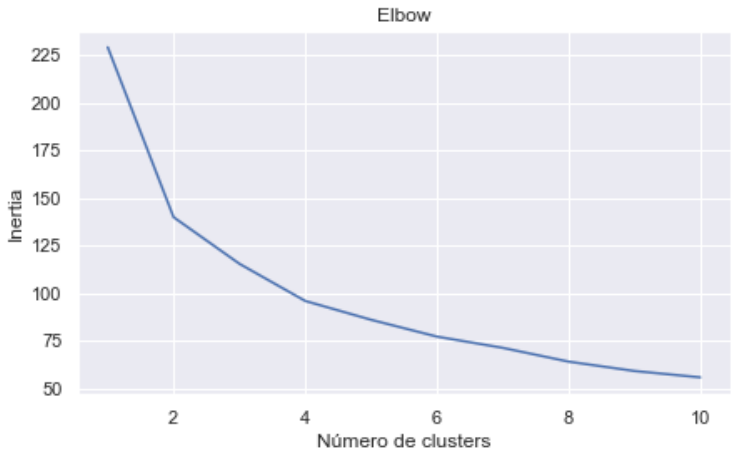
podría realizar las tareas relacionadas con los hijos fácilmente) que en un pareja en la que ambos trabajan.

- Por último, la variable sobre la que gira el caso de estudio *SATISFACENINOS* siendo una escala del 0 al 10, donde el 10 es el mayor grado de satisfacción.

Este caso de estudios está compuesto 1272 datos.

2.1. KMeans

Al ser el primer caso de estudio, la configuración con la que se va a aplicar *KMeans* es la predeterminada excepto el número de clusters, que se va a determinar mediante el *Elbow Method*

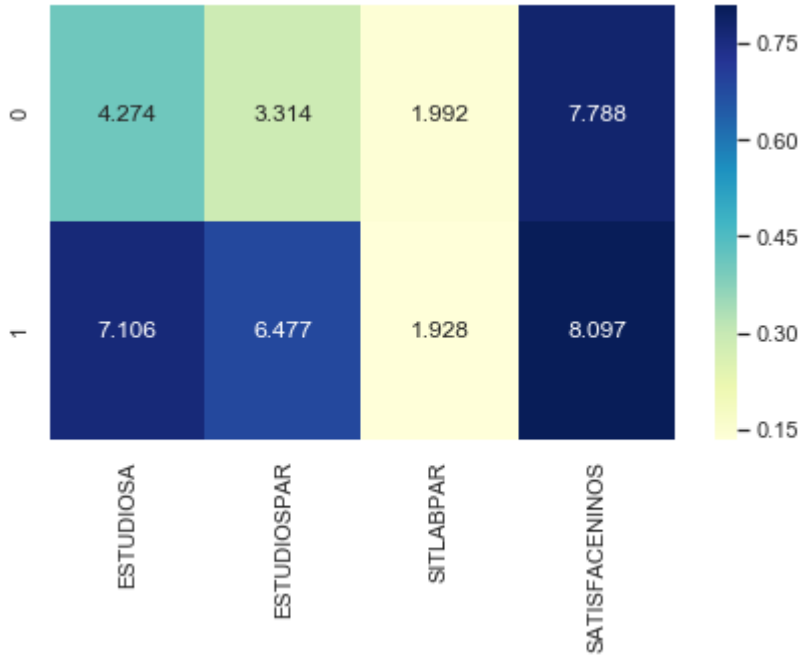


(b) Representación del método *Elbow* para el primer caso de estudio.

Nº clusters	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.01	0.36146	810.651	1: 688 (54.09 %) 0: 584 (45.91 %)
3	0.02	0.34902	624.660	1: 613 (48.19 %) 2: 485 (38.13 %) 0: 174 (13.68 %)
4	0.02	0.30854	587.479	3: 517 (40.64 %) 2: 326 (25.63 %) 1: 288 (22.64 %) 0: 141(11.08 %)
5	0.03	0.26229	525.493	3: 377 (29.64 %) 4: 292 (22.96 %) 2: 285 (22.41 %) 0: 212 (16.67 %) 1: 106 (8.33 %)
8	0.04	0.24859	452.881	0: 275 (21.62 %) 6: 229 (18.00 %) 1: 216 (16.98 %) 2: 193 (15.17 %) 5: 156 (12.26 %) 4: 139 (10.93 %) 3: 42 (3.30 %) 7: 22 (1.73 %)

A la vista de la representación 1(b), donde con k=2 empieza a decrementar linealmente, y de los resultados, se escoge la configuración con k=2.

Por tanto, con esta configuración se obtienen el siguiente Heatmap y Scattermatrix:



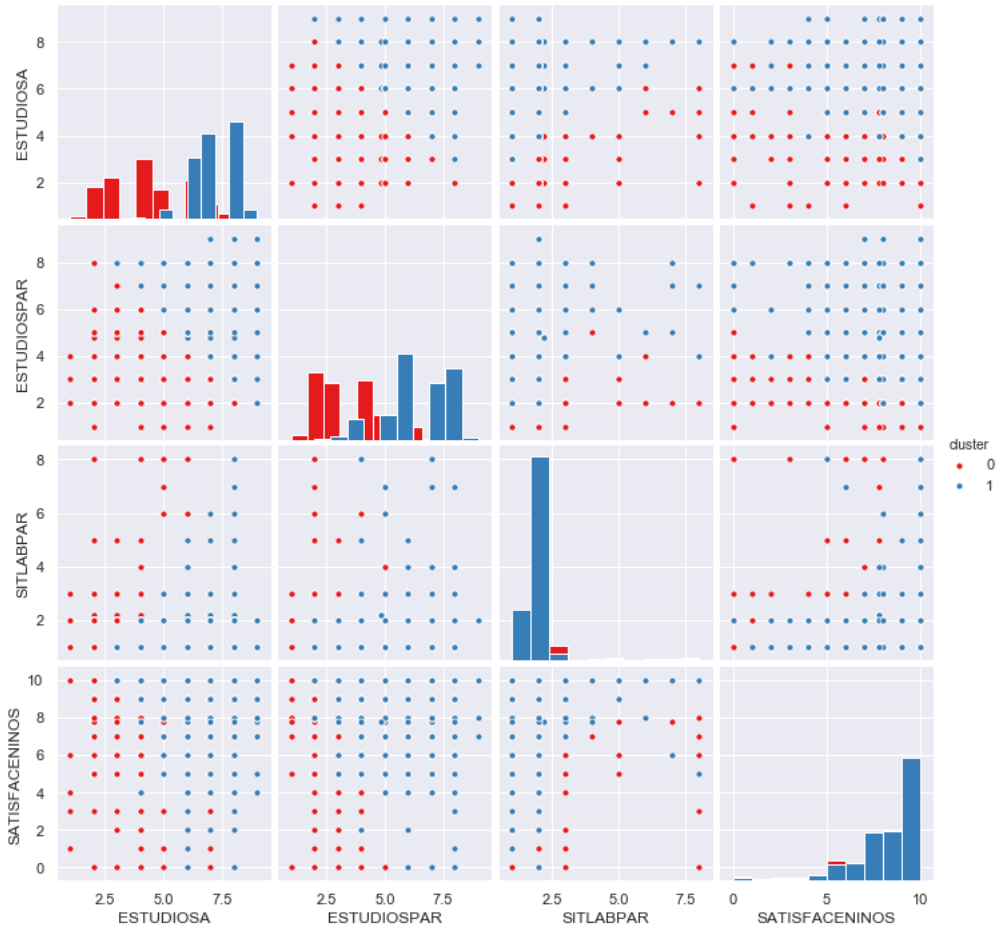
(c) Heatmap de KMeans del primer caso de estudio.

Como se puede ver en el heatmap 1(c), las características *SITLABPAR, SATISFACENINOS* son muy parecidas en ambos clusters. Mientras que *ESTUDIOSA, ESTUDIOSPAR* diferencian ambos clusters.

Referente a la variable *ESTUDIOSA*, en el cluster 0 se encuentra un grupo de mujeres cuyo nivel de estudios es 'Segunda etapa de educación secundaria y similar' y en el cluster 1 se encuentran el grupo de mujeres con un nivel de estudios 'Grados universitarios, diplomados universitarios, títulos propios universitarios de experto o especialista, y similares'.

La característica *ESTUDIOSPAR*, referente a los estudios alcanzados por la pareja, en el cluster 0 se encuentran los hombres con un nivel educativo de 'Primera etapa de educación secundaria y similar', mientras que en el cluster 1 se encuentran los hombres con un nivel educativo de 'Enseñanzas de formación profesional, artes plásticas y diseño y deportivas de grado superior y equivalentes'.

De este heatmap, podemos extraer como información que las mujeres, suelen tener un nivel educativo más alto que los hombres. A continuación, lo anteriormente mencionado se va a visualizar en el Scattermatrix.



(d) ScatterMatrix de KMeans del primer caso de estudio.

Lo anteriormente mencionado se puede ver en el scatter matrix 1(d), donde en *ESTUDIOSA*, *ESTUDIOSPAR y SATISFACENINOS* se ven los clusters bien diferenciados. En *ESTUDIOSA* el cluster 0 son las personas con un nivel de estudios menor que en el otro cluster. Esto pasa exactamente igual en la variable *ESTUDIOSPAR*, donde el cluster 0 tiene los individuos con menor nivel de estudios.

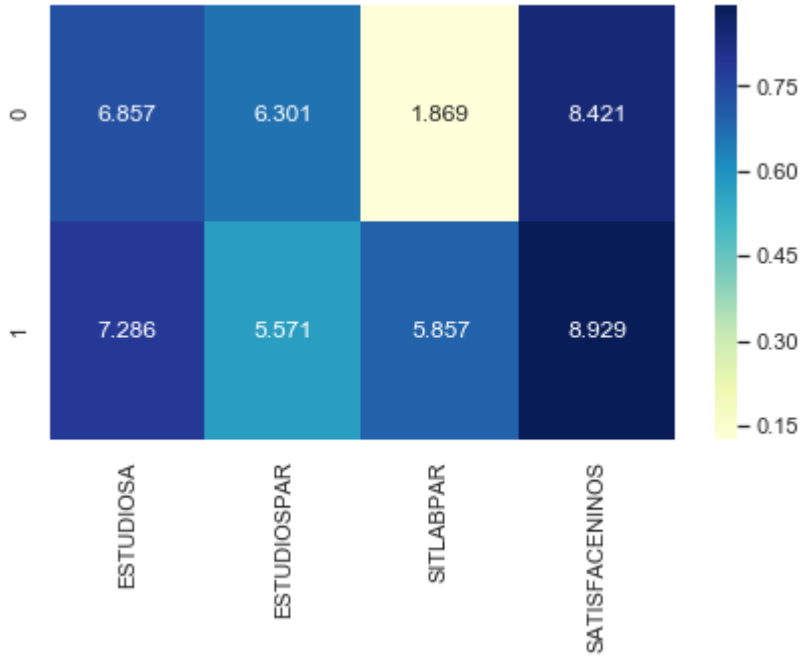
Por otro lado, con las características *SITLABPAR* y *SATISFACENINOS* no se puede extraer gran información de cómo están segmentados los grupos de individuos aunque, en *SATISFACENINOS*, parece que en el cluster 1 se agrupan las mujeres con un grado de satisfación mayor.

2.2. MeanShift

Meanshift estima, de forma autónoma, el número de clusters óptimo. A pesar de querer ver cómo se defendían las diferentes técnicas con su configuración básica, con meanshift he tenido que cambiar un parámetro para que devolviera un resultado mejor. El parámetro que he modificado es *min_ bin_ freq*. En la siguiente tabla se muestra los diferentes resultados cambiando ese parámetro:

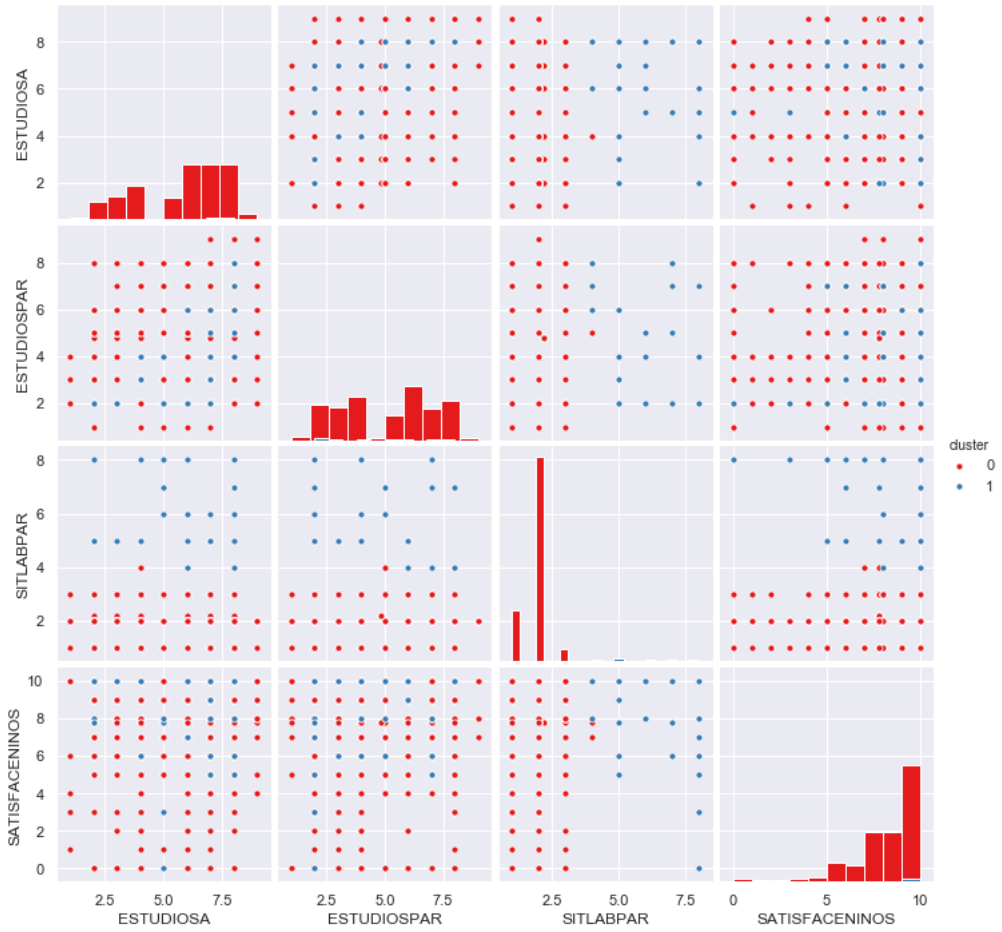
Nº clusters	Min_ bin_ freq	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	2	0.31	0.36864	63.716	0: 1242 (97.64 %) 1: 30 (2.36 %)
4	1	0.61	0.19870	49.899	0: 1218 (95.75 %) 2: 27 (2.12 %) 1: 15 (1.18 %) 0: 12(0.94 %)

A la vista de los resultados, se va a ejecutar con *min_ bin_ freq*=2 a pesar de que el valor de Calinski-Harabasz deje que desear. Además, con *min_ bin_ freq*=2, se generan dos clusters, número óptimo según *Elbow method*.



(e) Heatmap de Meanshift del primer caso de estudio.

Como se puede ver en heatmap 1(e), Meanshift ha segmentado de forma diferente que KMeans. *ESTUDIOSA, ESTUDIOSPAR* y *SATISFACENINOS* apenas diferencian ambos grupos mientras que la característica que más diferencia los grupos en este caso, al contrario que con KMeans, es *SITLABPAR*



(f) ScatterMatrix de Meanshift del primer caso de estudio.

Ya se había mencionado anteriormente, pero se puede volver a afirmar que la variable que distingue ambos grupos es *SITLABPAR*, viéndose claramente dos grupos, menor y mayor que el valor 5. El resto de variables tienen ambos clusters solapados y sin apenas distinción. Por tanto, considero que Meanshift no es el más adecuado para este caso de estudio.

2.3. DBSCAN

DBSCAN necesita dos parámetros para poder ejecutarse:

- Epsilon: La máxima distancia entre dos puntos para que se

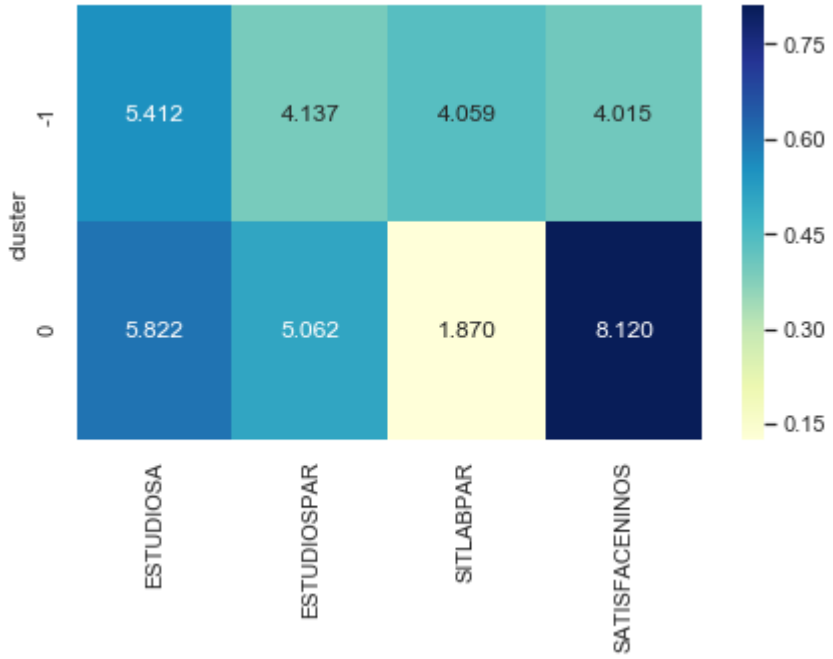
consideren del mismo cluster. Probablemente este sea el parámetro más importante para obtener una buena segmentación con DBSCAN.

- Min_samples: El número de punto necesarios para que un punto se considere centroide.

Según [8], el valor óptimo de *eps* es 0.3 y de *min_samples* es 5 pero vamos a realizar diferentes configuraciones para comprobar la veracidad del documento:

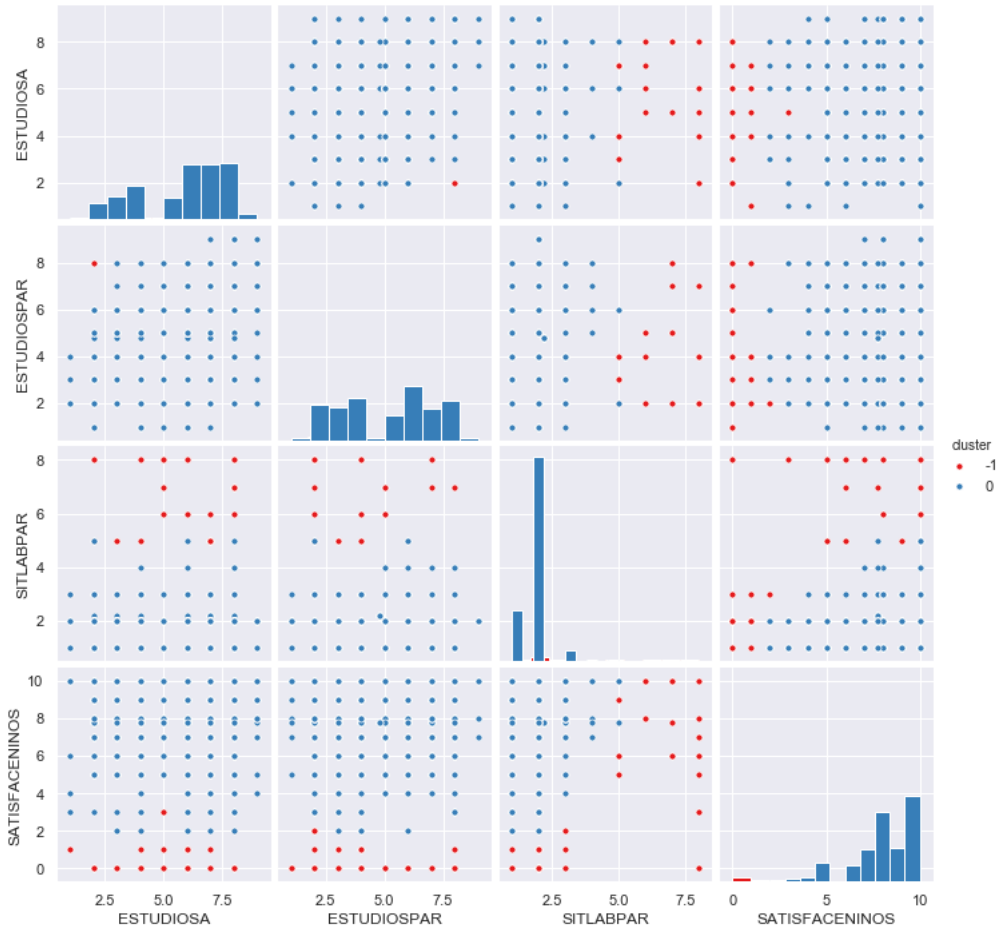
Nº clusters	Epsilon	Min_samples	T(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.3	5	0.04	0.45118	45.695	0: 1259 (98.98 %) -1: 13 (1.02 %)
3	0.2	5	0.03	0.33710	43.688	0: 1220 (95.91 %) -1: 46 (3.62 %) 1: 6 (0.47 %)
2	0.3	5	0.04	0.41866	61.119	0: 1245 (97.88 %) -1: 27 (2.12 %)
2	0.3	50	0.04	0.41110	81.475	0: 1221 (95.99 %) -1: 51 (4.01 %)

Si observamos los resultados, aparece un cluster con identificador -1, esto es debido a que en ese cluster se han agrupado los datos con ruido. La configuración con mejor Silhouette es la configuración aconsejada por [8] mientras que la configuración de *eps*=0.3 y *min_samples*=50 tiene mejor Calinski-Harabasz. En este caso voy a coger la última configuración mencionada.



(g) Heatmap de DBSCAN del primer caso de estudio.

Pese a ser un cluster formado por datos con ruido, la característica *SITLABPAR* y *SATISFACENINOS* están bien diferenciados en ambos clusters.



(h) ScatterMatrix de DBSCAN del primer caso de estudio.

Como he dicho anteriormente, la característica *SITLABPAR* divide de forma correcta el conjunto de datos aunque la mayoría de elementos formen parte del cluster 0. Esto me hace pensar que DBSCAN tampoco es una buena opción para este caso de estudio, habrá que explorar nuevas configuraciones en casos de estudios posteriores.

2.4. Birch

Antes de comenzar, he de decir que he tenido que modificar el úmbral porque si no, no lograba ejecutarse (solo encontraba un cluster). Por tanto, el úmbral va a ser 0.4

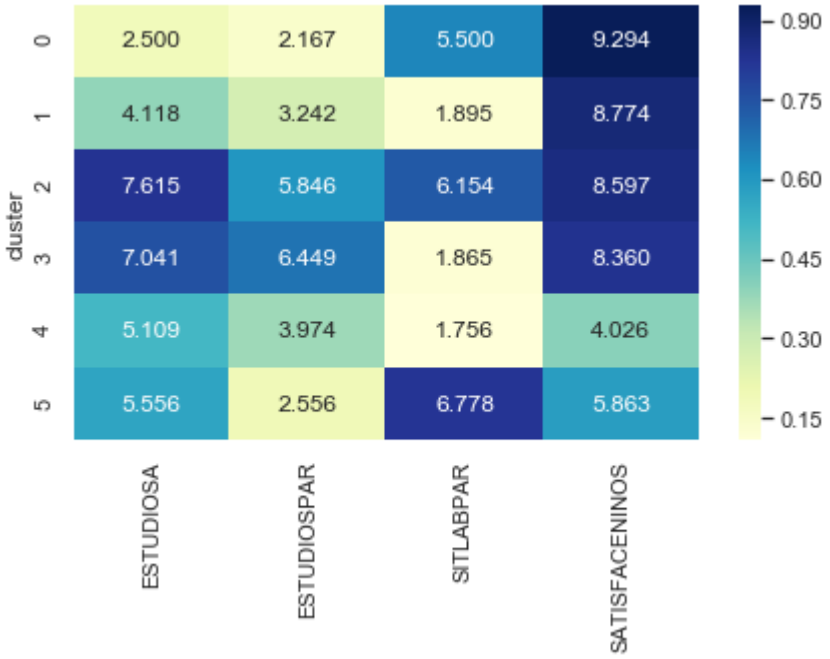
Birch necesita, como parámetro, el número de clusters. Por tanto, he probado las siguientes configuraciones:

Nº clusters	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.03	0.38192	64.392	1: 1244 (97.80 %) 0: 28 (2.20 %)
3	0.03	0.29357	38.279	0: 1244 (97.80 %) 1: 22 (1.73 %) 2: 6 (0.47 %)
4	0.03	0.31823	283.934	3: 665 (52.28 %) 0: 579 (45.52 %) 1: 22 (1.73 %) 2: 6(0.47 %)
5	0.15	0.34034	358.549	3: 665 (52.28 %) 1: 423 (33.25 %) 4: 156 (12.26 %) 0: 22 (1.73 %) 2: 6 (0.47 %)

A la vista de los resultados, voy a estudiar la configuración de 5 clusters [1] pues tiene un equilibrio mayor entre Silhouette y Calinski-Harabasz que la configuración de un dos clusters, aunque éste último tenga mejor Silhouette.

En todas las configuraciones hay un cluster (o varios) minoritarios. Considero que dichos clusters estarán formados por *outliers* que no se agrupan con el cluster mayoritario.

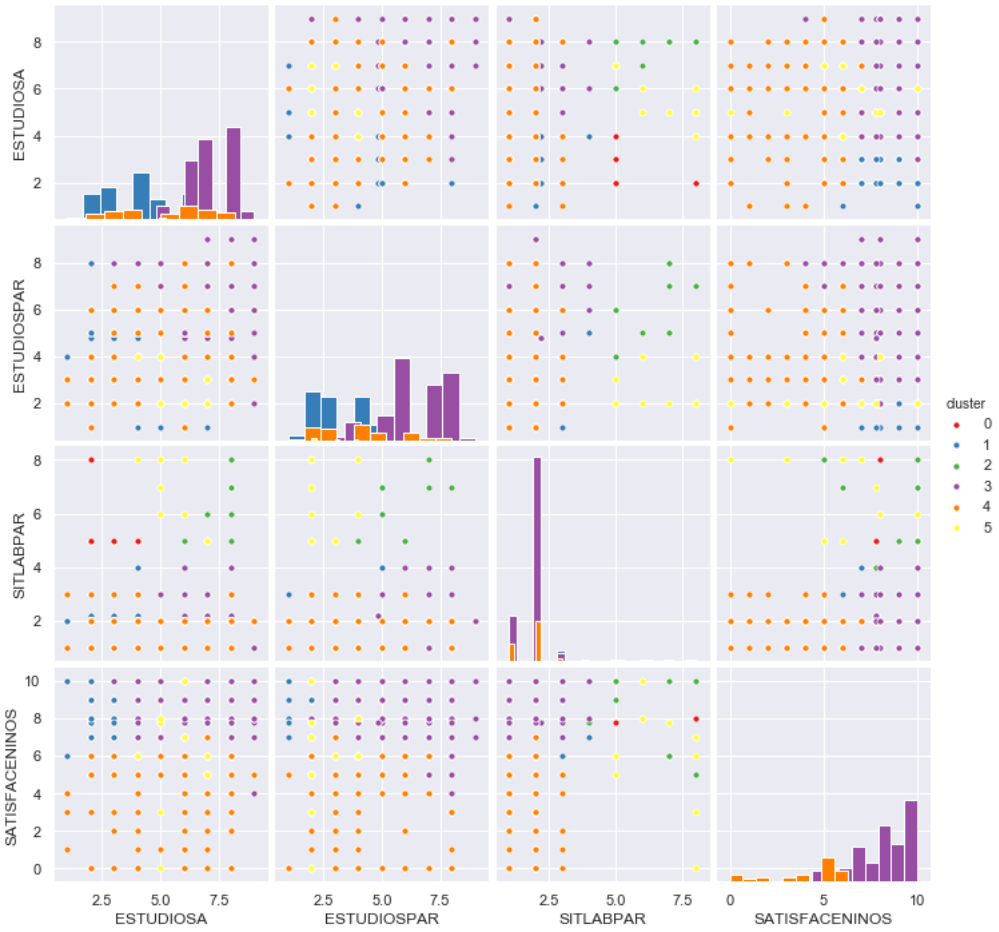
[1] A partir de 6 clusters, Silhouette empezaba a bajar de nuevo y por eso dejé de probar configuraciones alternativas.



(i) Heatmap de Birch del primer caso de estudio.

El cluster 0 agrupa a las mujeres y sus parejas con estudios menores pero que, a la vez, tienen un grado de satisfacción bastante alto. Por otro lado, está el cluster 2, que agrupa mujeres y sus parejas con un nivel alto educativo, un grado de satisfacción alto y la situación laboral de la pareja es 'Incapacidad para trabajar'. Mientras que el cluster 3, agrupa características semejantes que el cluster 2 pero la situación laboral de la pareja es que trabaja por su cuenta.

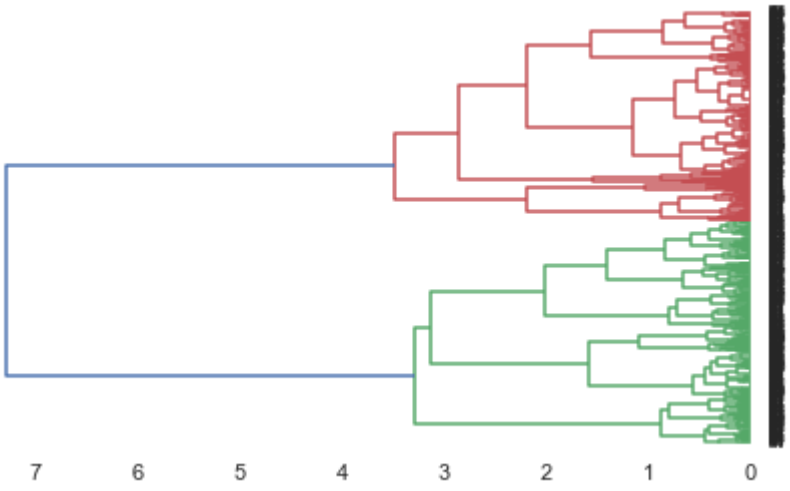
Todos los clusters tienen una característica diferente al resto, por ejemplo el cluster 3 y 2 que se diferencian en la situación laboral. O, por ejemplo, el cluster 1 y 4 que se diferencian en el grado de satisfacción. Aunque haya clusters con pocos elementos y en el scatter matrix es difícil distinguir los clusters más pequeños, cada cluster tiene una característica peculiar respecto al resto.



(j) ScatterMatrix de Birch del primer caso de estudio.

2.5. Aglomerativo

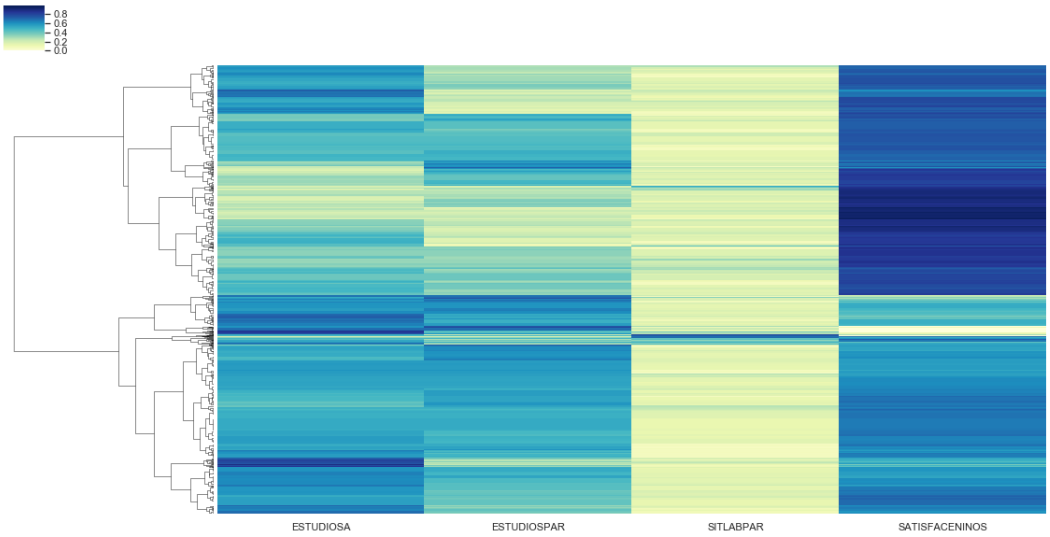
Para determinar el número de clusters de esta técnica se va a hacer uso del dendograma:



(k) Dendograma del primer caso de estudio.

En vistas del dendograma 1(k), se van a elegir dos clusters.

Nº clusters	Tiempo(s)	Tamaño de cada cluster
2	0.03	1: 513 (51.30 %) 0: 487 (48.70 %)



(l) Clustermap del primer caso de estudio.

Se puede ver, claramente, que en un cluster el grado de satisfacción, *SATISFACENINOS*, es, generalmente, bastante mayor que en el otro cluster. Además, en el cluster donde el grado de satisfacción es mayor, los estudios de la pareja son inferiores a los del otro cluster. Por último, pasa exactamente lo mismo con el nivel de estudios de la mujer, se observa un nivel inferior de estudios donde el grado de satisfacción es mayor.

2.6. Interpretación de la segmentación

A la luz de los resultados de las diferentes técnicas, se puede concluir que las mujeres con mayor grado de satisfacción son las que tienen un nivel de estudios menor. Si nos fijamos en la variable *SATISFACENINOS* comparándola con *SITLABPAR* del scatter matrix 1(d), hay una laguna por la parte superior izquierda de la gráfica, correspondiéndose con valores bajos de satisfacción. Esto puede ser debido a que los estudiantes, valor 4 de la característica, al ser jóvenes, están más concienciados con el reparto de tareas. Además, los valores 5, 6 y 7, jubilados, personas incapacitadas para trabajar y personas dedicadas a las labores del hogar respectivamente, tienen también una alta tasa de satisfacción. Por otro lado, existen muchas mujeres que están descontentas pese a que su pareja está

desempleada actualmente ('Parado' es el valor 3 de *SITLABPAR*). Desconozco cuáles pueden ser los motivos del descontento pero las mujeres con parejas paradas deberían tener un grado de satisfacción alto, al igual que con los jubilados o personas incapacitadas para trabajar, por ejemplo.

Como conclusión, las mujeres con menor nivel educativo están más contentas con el reparto de tareas y que, por regla general, hay casos de satisfación y de insatisfación, por igual, en mujeres con parejas trabajando siendo curioso que existan casos de mujeres descontentas con parejas en estado de desempleo.

3. Caso de estudio 2: Estudios alcanzados y campos en los que se han especializado las mujeres.

```
#seleccionar casos
subset = censo.loc[((censo['EDAD']>25) & (censo['EDAD']<=40))]

#seleccionar variables de interés para clustering
usadas = ['CAMPOA', 'ESTUDIOSA', 'INGREHOG' ,'ESTUDIOSPAR', 'CAMPOPAR']
```

(a) Código asociado al segundo caso de estudio

Antes de comenzar a tratar el tercer[2] caso de estudio con las diferentes técnicas, se muestra el código correspondiente al caso de estudio y una explicación de por qué se han elegido esas variables. La variable escogida para establecer el subconjunto de datos es la siguiente:

- A través de la edad, se eligen las mujeres entre 25 y 40 años.

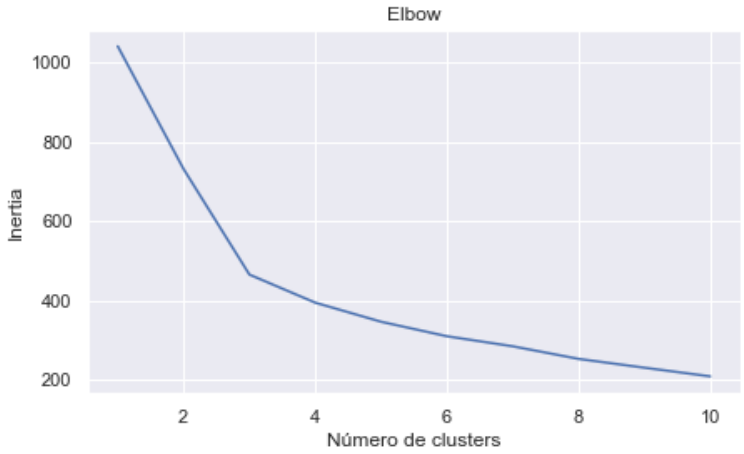
Por otro lado, las variables elegidas para realizar la segmentación son las siguientes:

- Tanto *ESTUDIOSA* y *ESTUDIOSPAR* son variables elegidas para ver si hay algún tipo de relación entre el nivel de estudios alcanzados por ambas personas. Esta variable va desde *Menos que primaria* (código 1) y *Enseñanzas de doctorado* (código 9).
- *CAMPOA* y *CAMPOPAR* son características elegidas, al igual que las anteriores, para comprobar si hay alguna relación entre el campo de ambas personas de la pareja.
- Por último, la característica *INGREHOG* para visualizar qué hogares tienen más ingresos mensuales dependiendo del campo de la mujer y su pareja.

Este caso de estudios está compuesto 5097 datos.

[2]Mi idea inicial era realizar un estudio sobre mujeres científicas/ingenieras informáticas en toda España, el problema vino cuando el subconjunto tenía un total de 240 mujeres, aproximadamente, siendo un subconjunto muy pequeño para poder realizar segmentación de forma correcta.

3.1. KMeans



(b) Representación del método *Elbow* para el segundo caso de estudio.

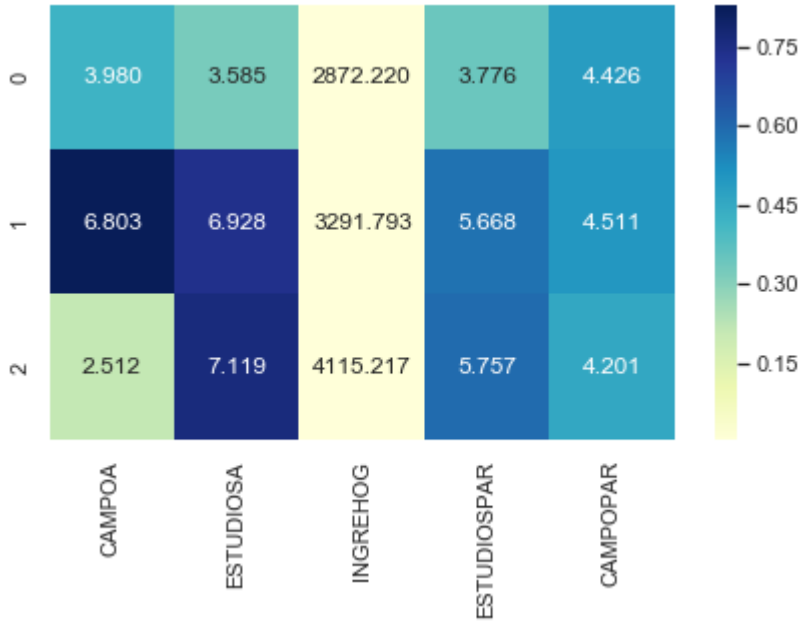
Nº clusters	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.02	0.29892	2153.332	0: 3062 (60.07 %) 1: 2035 (39.93 %)
3	0.04	0.40064	3145.320	1: 1997 (39.18 %) 2: 1942 (38.10 %) 0: 1158 (22.72 %)
4	0.04	0.37475	2775.554	0: 1887 (37.02 %) 1: 1307 (25.64 %) 2: 1084 (21.27 %) 3: 819 (16.07 %)
6	0.07	0.33807	2405.878	0: 1224 (24.01 %) 5: 1200 (23.54 %) 1: 880 (17.27 %) 4: 777 (15.24 %) 2: 657 (12.89 %) 3: 359 (7.04 %)

A la vista de la representación 1(c), donde decrece de forma rápida hasta k=3 y, a partir de ahí, va decreciendo de forma lineal, y de los resultados, se escoge la configuración con k=3. Además del número de clusters, se ha modificado el parámetro *max_ iters* a un valor de 500, 200 más que en su versión predeterminada.

```
k_means = KMeans(init='k-means++', n_clusters=num_clust, n_init=5, max_iter=500)
```

(c) Configuración del algoritmo KMeans para el segundo caso de estudio.

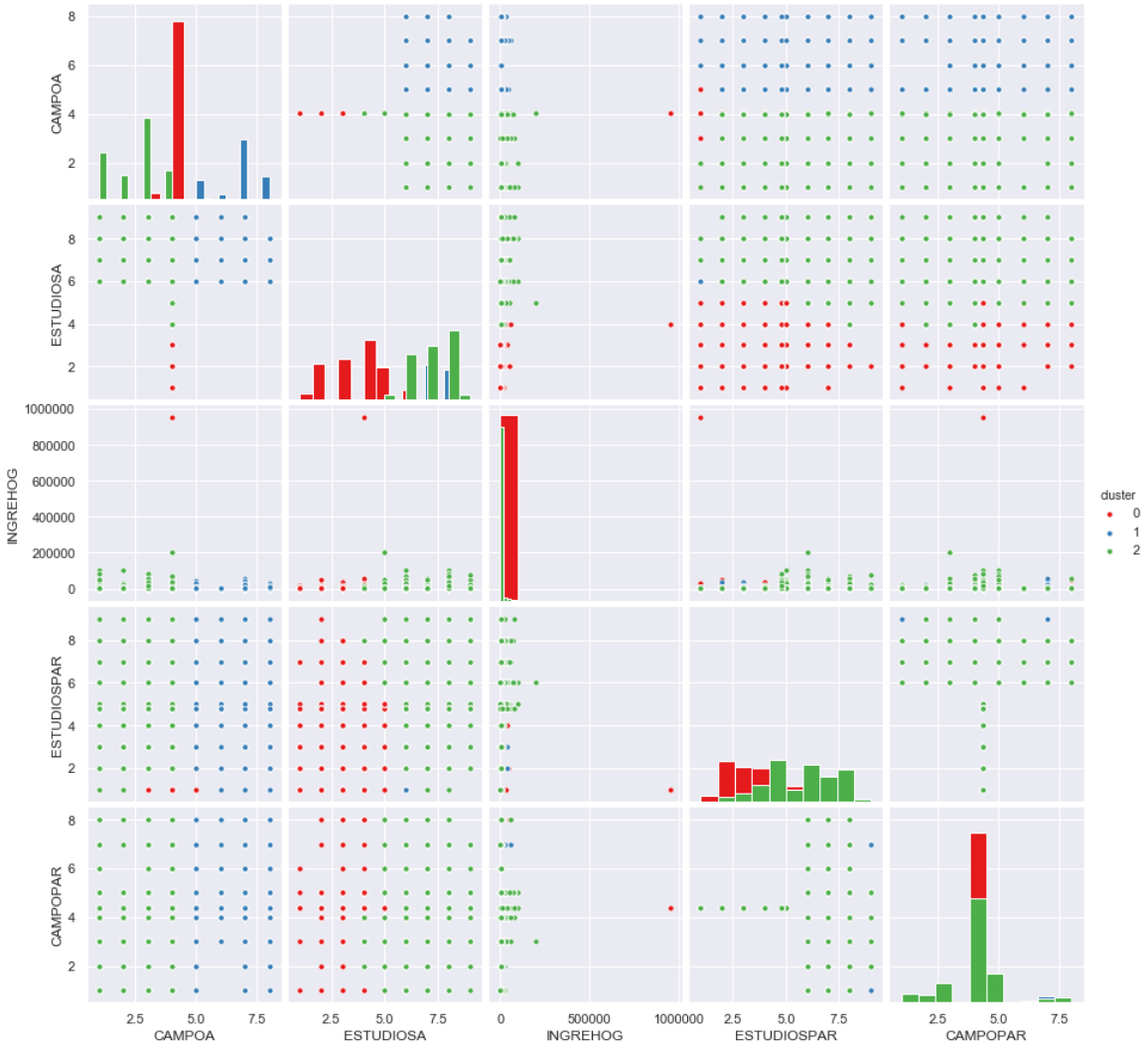
Por tanto, con esta configuración se obtienen el siguiente Heatmap y Scattermatrix:



(d) Heatmap de KMeans del segundo caso de estudio.

Como se puede observar en la característica *CAMPOA*, los 3 clusters están bien diferenciados, agrupándose en el cluster 0 las mujeres científicas/ingenieras informáticas, en el cluster 1 médicas, enfermeras, farmacéuticas, trabajadoras sociales, etc. Y, por último, el cluster 2 agrupa las mujeres especializadas en artes y humanidades. Los estudios alcanzados por las mujeres médicas y artistas son grados universitarios mientras que, las mujeres científicas o ingenieras, sus estudios alcanzados llegan al nivel de 'Primera etapa de educación secundaria y similar', siendo esto curioso pues necesitas estudios universitarios para dichos empleos.

Si nos fijamos en los 3 clusters se han agrupado las mujeres con hombres que son científicos o ingenieros informáticos.



(e) ScatterMatrix de KMeans del segundo caso de estudio.

Analizando el heatmap hemos observado que las mujeres científicas/ingenieras informáticas tienen un nivel de estudios alcanzado inferior al resto de mujeres, si nos fijamos en la característica *ESTUDIOSA* comparándola con *CAMPOA*, todos los campos tienen un nivel de 'Grados universitarios de hasta 240 créditos ECTS' excepto el campo de ciencias e ingenieria informática. Debido a que para obtener dichos empleos necesitas un título universitario, considero que los datos con un nivel de estudios inferior a un título universitario y

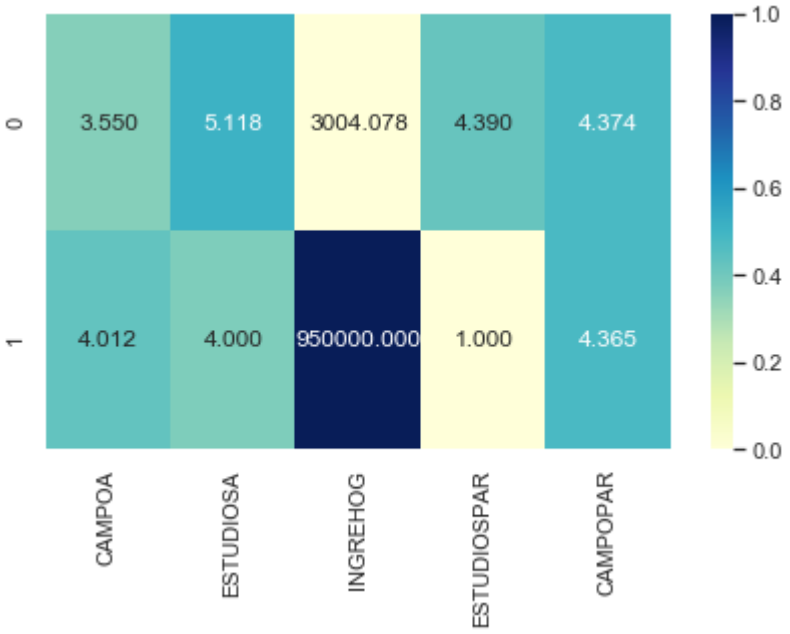
un campo de estudios de ciencias e ingenieria informática, son datos ruidosos.

Por otro lado, todos los clusters en la característica *INGREHOG* están agrupados en ingresos de menor a 5000 euros mensuales pero hay un elemento del cluster 0 que asegura tener ingresos mensuales de un millón de euros. Por tanto, lo considero un dato con ruido.

3.2. MeanShift

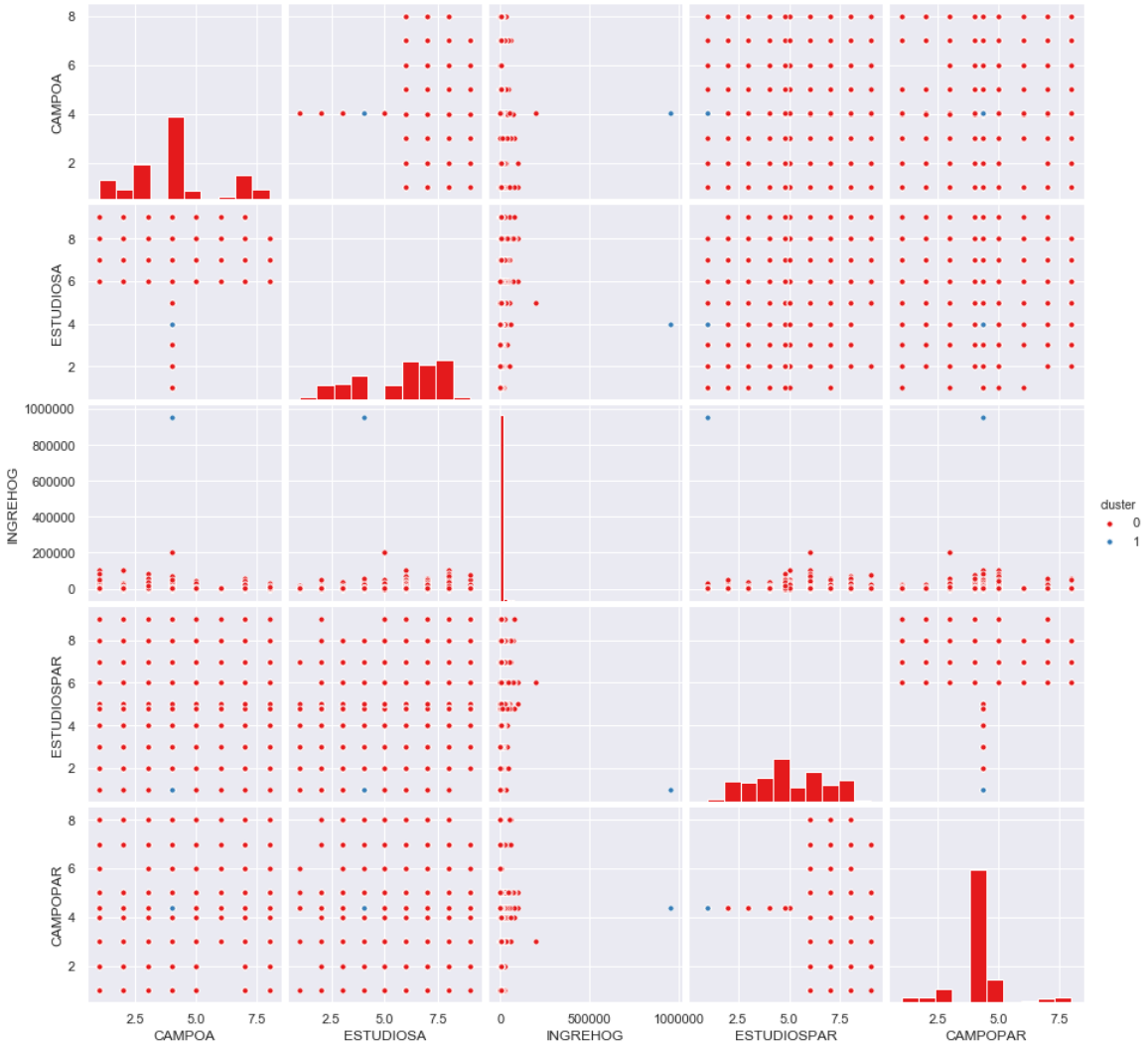
Con los valores predeterminados [3], los resultados son los siguientes:

Nº clusters	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.04	0.51442	6.291	0: 5096 (99.98 %) 0: 1 (0.02 %)



(f) Heatmap de Meanshift del segundo caso de estudio.

[3] Apenas he podido tocar parámetros con Meanshift puesto que, intentaba modificar *min_bin_freq* o *bandwidth*, solo lograba sacar un cluster y, por tanto, daba error y no se podía ejecutar.



(g) ScatterMatrix de Meanshift del segundo caso de estudio.

Como ya se apreciaba en la segmentación realizada por KMeans, en la característica *INGREHOG* había un dato con ruido que decía que los ingresos eran de un millón de euros. Dicho elemento lo ha considerado Meanshift como un único cluster.

No se ha podido extraer mejor segmentación que la obtenida pues no permitía ejecutarlo con otra configuración.

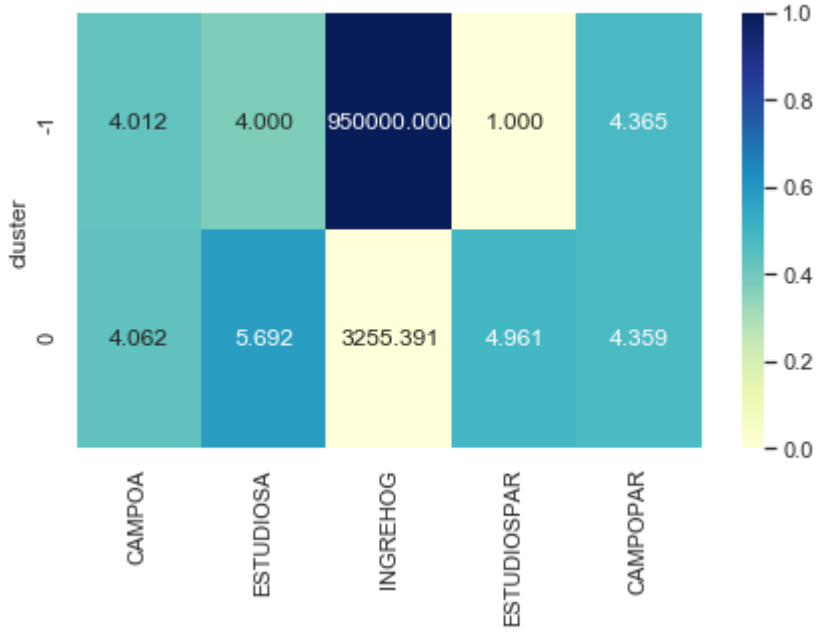
3.3. DBSCAN

Se ha probado las siguientes configuraciones:

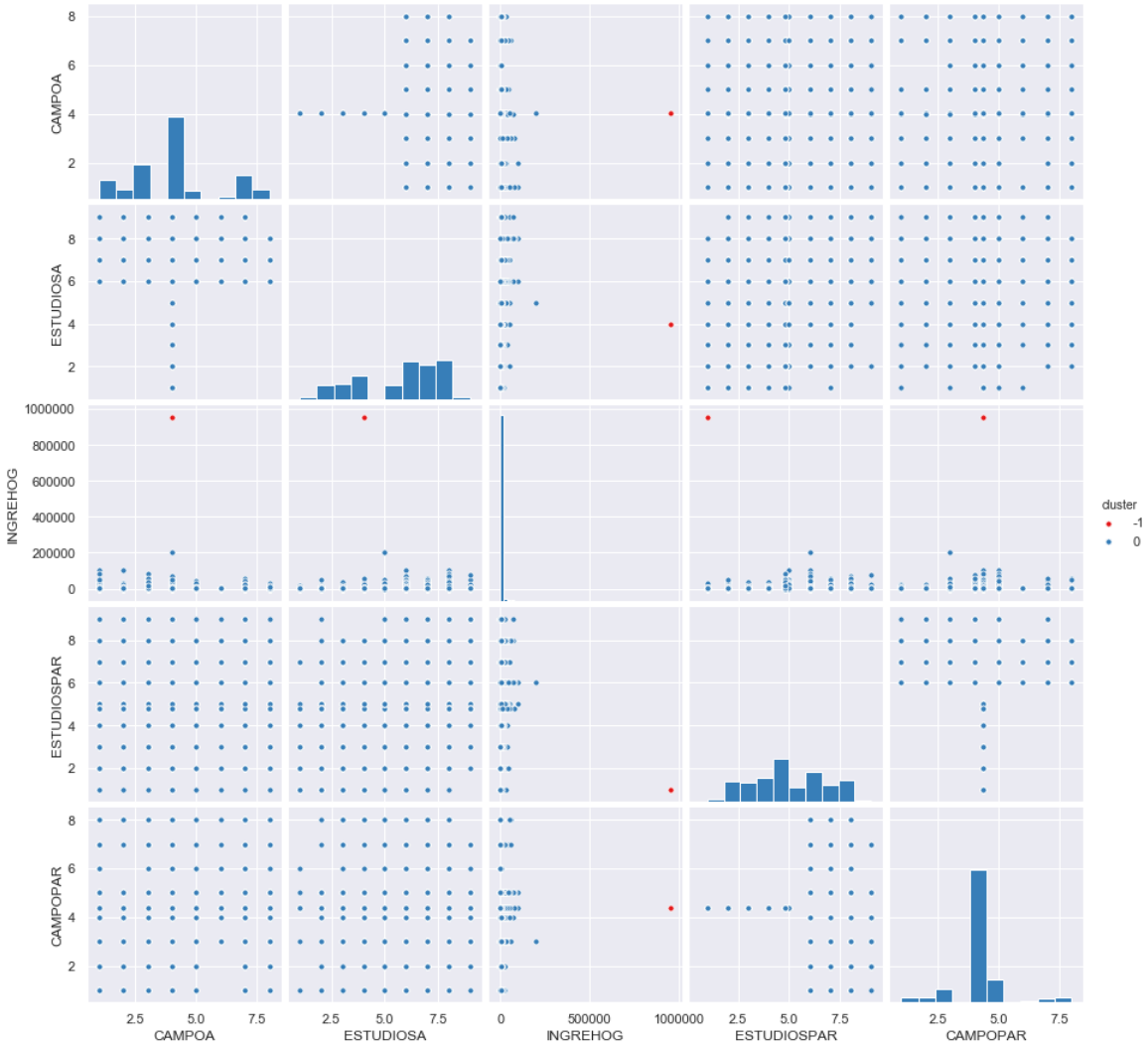
Nº clusters	Epsilon	Min_samples	T(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.3	10	0.052	0.51442	6.291	0: 5096 (99.98 %) -1: 1 (0.02 %)
2	0.2	10	0.28	0.31370	5.627	0: 5093 (99.92 %) -1: 4 (0.08 %)
2	0.2	10	0.28	0.31370	5.627	0: 5093 (99.92 %) -1: 4 (0.08 %)
104	0.1	10	0.13	0.49604	93.203	Más de 100 clusters.

La configuración con mejor Silhouette es la configuración *eps* = 0.3 y *min_samples* = 10 pero la medida Calinski-Harabasz es muy malo. La configuración que balancea el coeficiente de Silhouette y Calinski-Harabasz es la configuración con *eps* = 0.1 y *min_samples* = 10 pero contiene más de 100 clusters diferentes y, tanto el heatmap como el scatter matrix, son prácticamente inapreciables.

Las gráficas obtenidas con *eps* = 0.3 y *min_samples* = 10 son:



(h) Heatmap de DBSCAN del segundo caso de estudio.



(i) ScatterMatrix de DBSCAN del segundo caso de estudio.

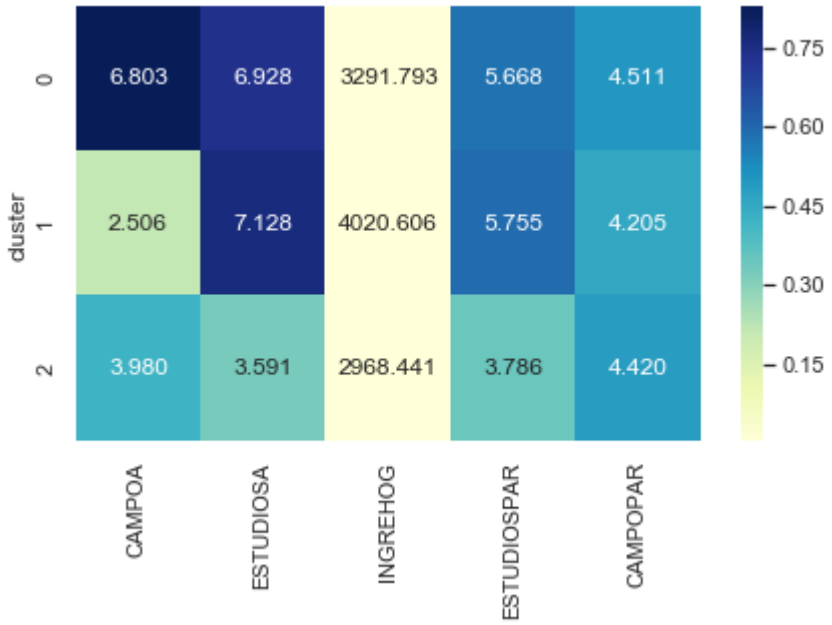
Vuelve a suceder el mismo fenómeno que con Meanshift, el dato con ruido se agrupa en un único cluster mientras que el resto se agrupa todo en el mismo cluster.

3.4. Birch

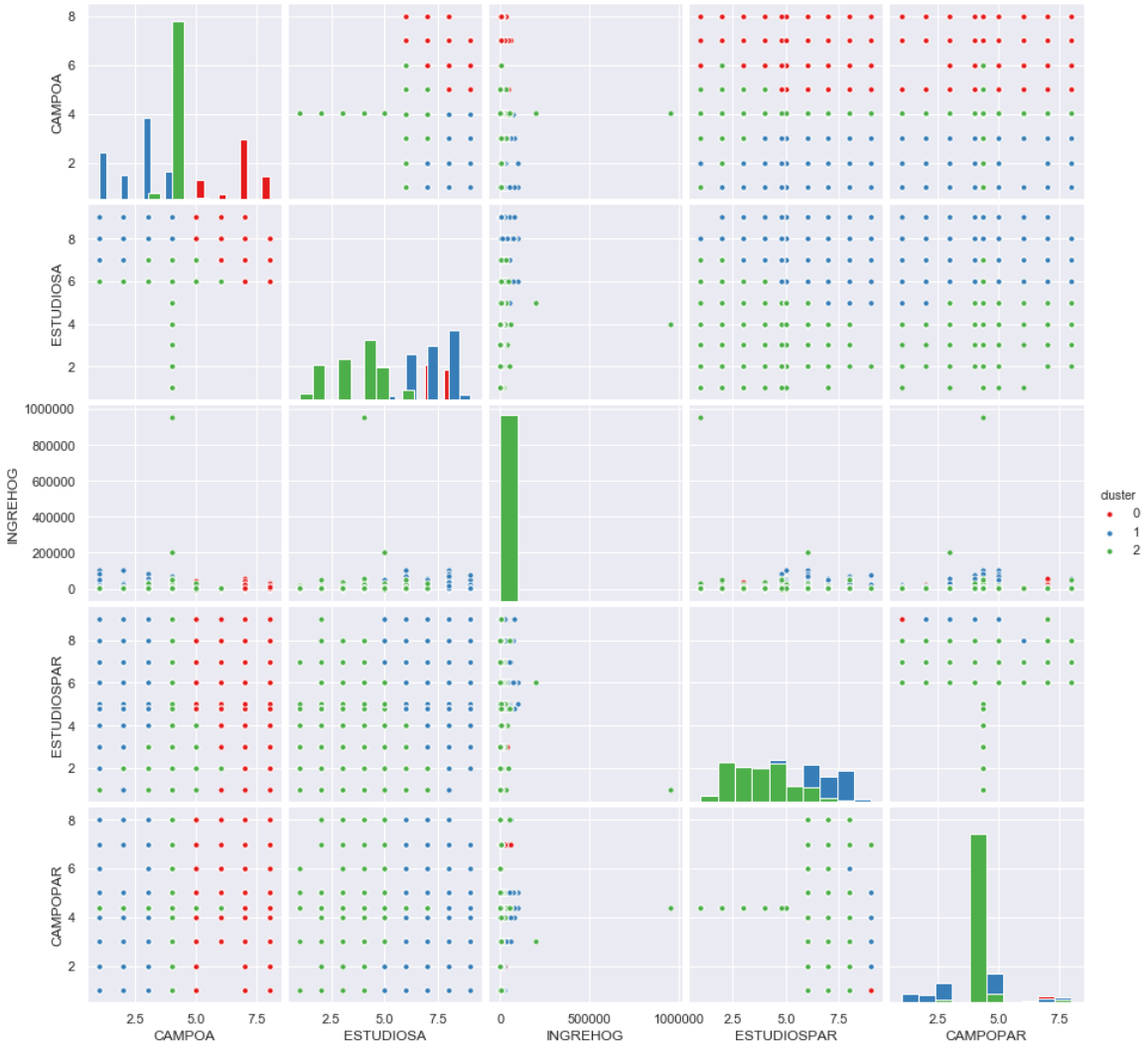
Se han probado las siguientes configuraciones:

Nº clusters	Umbral	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.2	0.10	0.33685	1753.329	0: 4146 (81.34 %) 1: 951 (18.66 %)
2	0.3	0.09	0.33358	1817.665	0: 4027 (79.01 %) 1: 1070 (20.99 %)
3	0.2	0.10	0.38254	2872.383	0: 2244 (44.03 %) 2: 1902 (37.32 %) 1: 951 (18.66 %)
3	0.3	0.10	0.39179	3006.559	2: 2033 (39.89 %) 0: 1994 (39.12 %) 1: 1070 (20.99 %)
3	0.4	0.12	0.40089	3145.029	2: 2005 (39.34 %) 1: 1934 (37.94 %) 0: 1158 (22.72 %)

A la vista de los resultados, se elige la configuración de k=3 y umbral = 0.4.



(j) Heatmap de Birch del segundo caso de estudio.

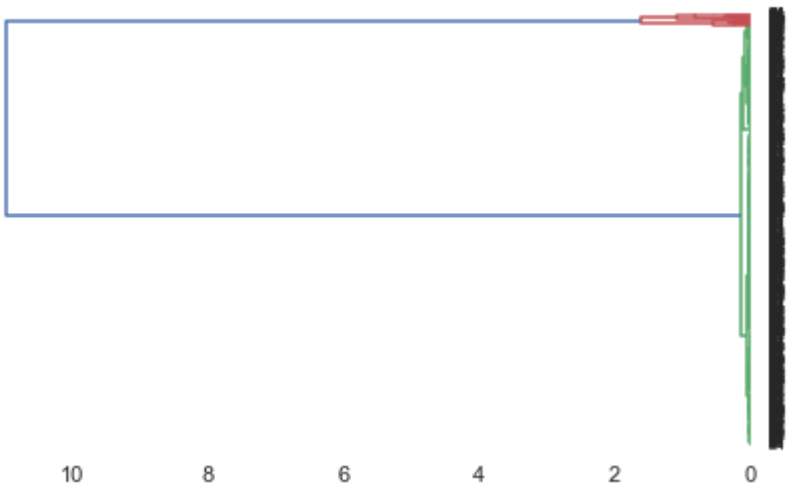


(k) ScatterMatrix de Birch del segundo caso de estudio.

Se logra una segmentación muy similar a la lograda por *KMeans*. Si nos fijamos en los heatmaps de ambas técnicas, son prácticamente iguales.

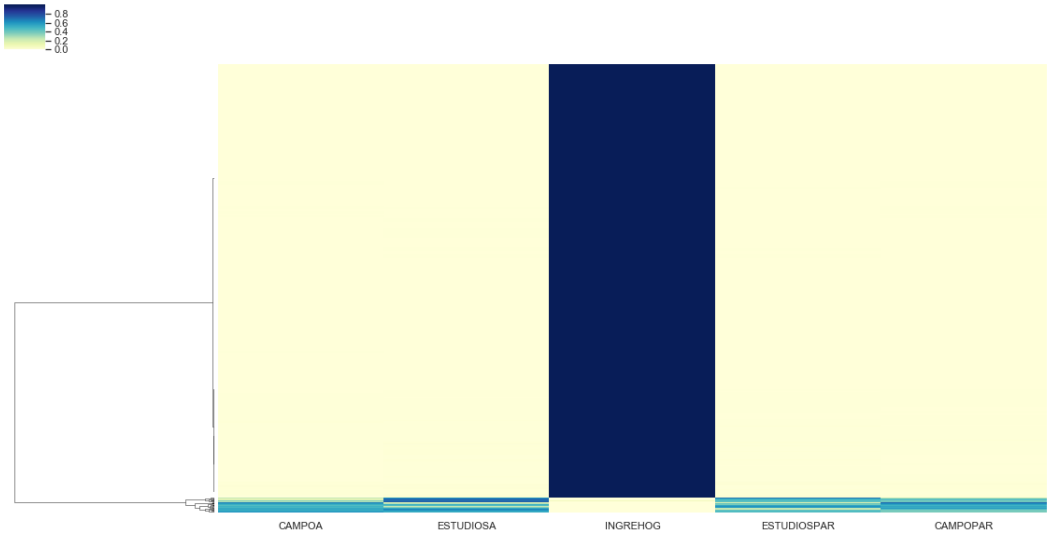
3.5. Aglomerativo

Para determinar el número de clusters de esta técnica se va a hacer uso del dendograma:



(l) Dendograma del segundo caso de estudio.

Nº clusters	Tiempo(s)	Tamaño de cada cluster
2	0.04	1: 967 (96.70 %) 0: 33 (3.30 %)



(m) Clustermap del segundo caso de estudio.

Se puede observar que, el dato ruidoso del que llevamos hablando todo este caso de estudio, ha fastidiado esta segmentación. Asigné dos clusters; en el cluster 1 está la mayoría de elementos, mientras que en el cluster 2 están los datos ruidosos. Probé a asignar más clusters por si el cluster mayoritario empezaba a fragmentarse pero, en su lugar, se dividía el cluster 0.

Además, he probado con las cuatro diferentes estrategias por si el resultado cambiaba pero no ha sido posible.

```
ward = AgglomerativeClustering(n_clusters=num_clusters, linkage='ward')
```

(n) Estrategia ward.

```
ward = AgglomerativeClustering(n_clusters=num_clusters, linkage='average')
```

(ñ) Estrategia average.

```
ward = AgglomerativeClustering(n_clusters=num_clusters, linkage='complete')
```

(o) Estrategia complete.

```
ward = AgglomerativeClustering(n_clusters=num_clusters, linkage='single')
```

(p) Estrategia single.

3.6. Interpretación de la segmentación

Tras los resultados obtenidos, se puede afirmar que los estudios alcanzados por las mujeres médicas y artistas son superiores que los alcanzados por las mujeres científicas. Además, prácticamente todas las mujeres tienen el mismo nivel de ingresos mensuales en el hogar, pese a los diversidad de campos, excepto las mujeres abogadas y científicas, que destacan por tener más ingresos en el hogar y las mujeres especializadas en la agricultura o turismo destacan por un nivel de ingresos menor. A priori, no hay ninguna relación entre el campo de estudio de la mujer y el de su pareja. Además, se repite el mismo fenómeno que con las mujeres, los hombres médicos alcanzan un nivel de estudio superior a los hombres científicos o ingenieros, aunque me temo que estos datos son ruidosos.

Pese a no formar parte del caso de estudio, existen muy pocas mujeres científicas e ingenieras informáticas, pues el conjunto de datos está compuesto por 5097 datos pero el subconjunto formado por las mujeres científicas o ingenieras, estaba formado por 240 datos, apróximadamente. Lo que me hace concluir que las mujeres siguen sufriendo un estigma social a la hora de decidir estudiar grados universitarios relacionados con ciencias o ingenierias.

4. Caso de estudio 3: Resentimiento de las mujeres por no haber tenido hijos

```
subset = censo.loc[((censo['EDAD']>=35) & (censo['DESEOHIJOS']==1))]
#subset = subset.loc[(censo['COD_CCAA']=='01')]

#seleccionar variables de interés para clustering
usadas = ['EDAD','EDADIDEAL', 'TEMPRELA', 'INGRESOS', 'RELIGION']
```

(a) Código asociado al tercer caso de estudio

Antes de comenzar a tratar el tercer caso de estudio con las diferentes técnicas, se muestra el código correspondiente al caso de estudio y una explicación de por qué se han elegido esas variables. Las variables escogidas para establecer el subconjunto de datos son las siguientes::

- A través de la edad, se eligen mujeres con más de 35 años.
- Con la característica *DESEOHIJOS* nos quedamos con las mujeres que se arrepienten por no haber tenido hijos o por no haber tenido las que les hubiera gustado tener.

Por otro lado, las variables elegidas para realizar la segmentación son las siguientes:

- Tanto *EDAD* y *EDADIDEAL* son variables elegidas para ver si hay algún tipo de relación la edad de la mujer y la edad que considera que es ideal.
- *TEMPRELA* es una característica que representa el tiempo de relación que lleva con pareja. La intención con esta variable es comprobar si, cuanto más lleva una persona en una relación, cambia su percepción de la edad ideal. *INGRESOS* es una característica que representa los ingresos mensuales de la mujer.

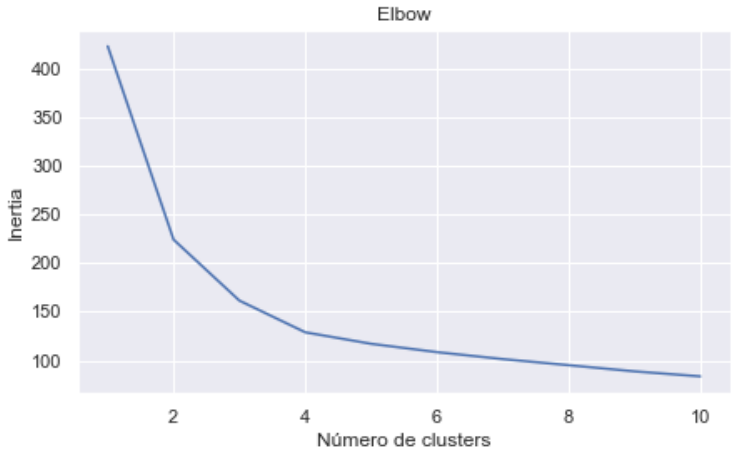
Se ha elegido para comprobar si se pueden suponer motivos económicos por el que la mujer no ha tenido hijos o no tanto como quería.

- Por último, la característica *RELIGION* para visualizar relaciones entre la religión y la edad ideal de tener hijos consideradas por personas de diferentes religiones.

Con estas variables, el caso de estudio quiere tratar cuáles son las edades ideales para tener hijos según mujeres que se resienten de no haber tenido hijos o tantos como les hubiera gustado, además se pretende ver si hay características, como la religión, que influyen en esa edad ideal.

Este caso de estudios está compuesto 1144 datos.

4.1. KMeans

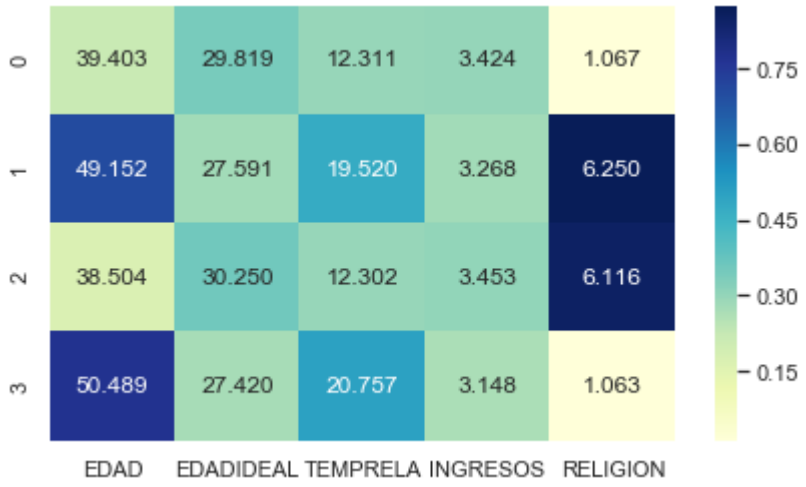


(b) Representación del método *Elbow* para el tercer caso de estudio.

Nº clusters	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.01	0.44509	1012.039	0: 745 (61.63 %) 1: 439 (38.37 %)
3	0.02	0.38218	922.782	1: 433 (37.85 %) 2: 390 (34.09 %) 0: 321 (28.06 %)
4	0.03	0.35777	865.698	0: 387 (33.83 %) 1: 317 (27.71 %) 3: 276 (24.13 %) 2: 164 (14.34 %)

Pese a que los mejores resultados son con la configuración de dos clusters, el estudio se va a realizar con la opción de cuatro clusters, pues la segmentación está mejor realizada pese a tener un coeficiente de Silhouette inferior.

Por tanto, con esta configuración se obtienen el siguiente Heatmap y Scattermatrix:



(c) Heatmap de KMeans del tercer caso de estudio.

(d) ScatterMatrix de KMeans del tercer caso de estudio.

Si nos fijamos en la variable *EDAD* relacionada con *RELIGION*, se puede ver claramente la segmentación:

- El cluster 0 agrupa las mujeres menores de 45 años que son católicas, protestantes o musulmanas.
- El cluster 1 agrupa las mujeres mayores de 45 años que son ateos, siguen otra religión de las especificadas o ha preferido no responder.

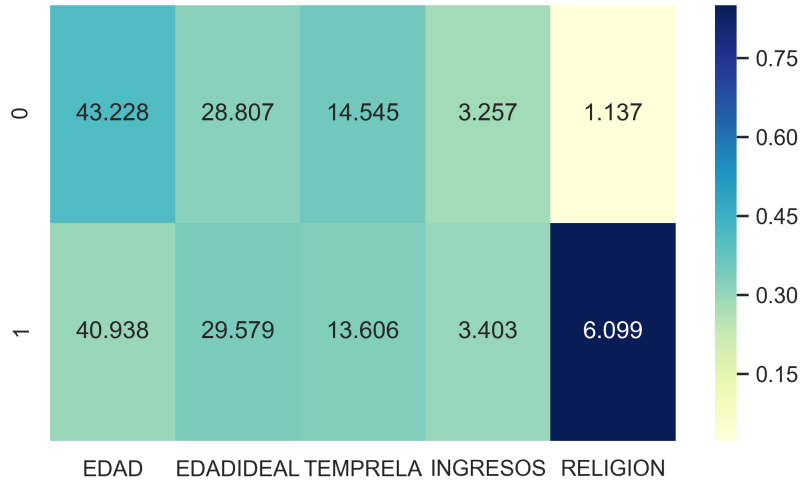
- El cluster 2 agrupa las mujeres menores de 45 años que son ateos, siguen otra religión de las especificadas o ha preferido no responder.
- El cluster 3 agrupa las mujeres mayores de 45 años que son católicas, protestantes o musulmanas.

4.2. MeanShift

La ejecución de Meanshift ha sido con los valores predeterminados y se ha obtenido este resultado:

Nº clusters	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.59	0.44524	1011.980	0: 707 (61.80 %) 0: 437 (38.20 %)

Meanshift con los valores predeterminados ha realizado la segmentación con $k = 2$, número óptimo de clusters según Elbow Method 1(b)



(e) Heatmap de Meanshift del tercer caso de estudio.

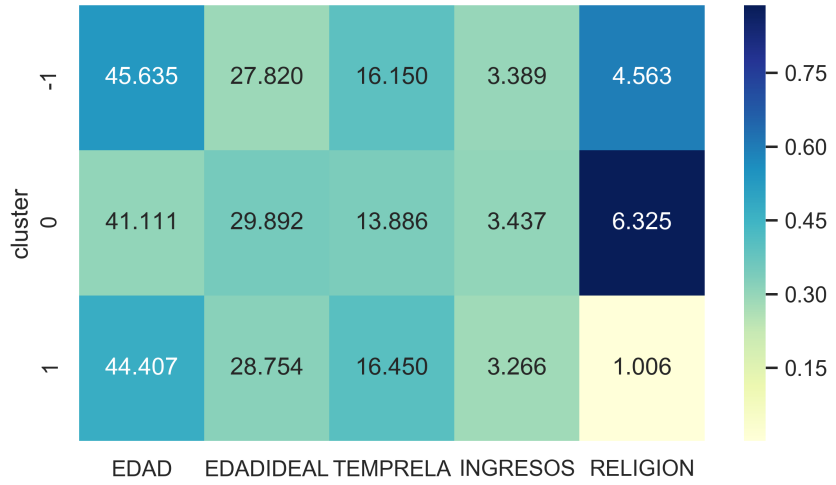
Ambos clusters son muy parecidos excepto en la característica *RELIGION* donde en el cluster 0 se agrupan las mujeres católicas y en el cluster 1 las mujeres ateas.

(f) ScatterMatrix de Meanshift del tercer caso de estudio.

4.3. DBSCAN

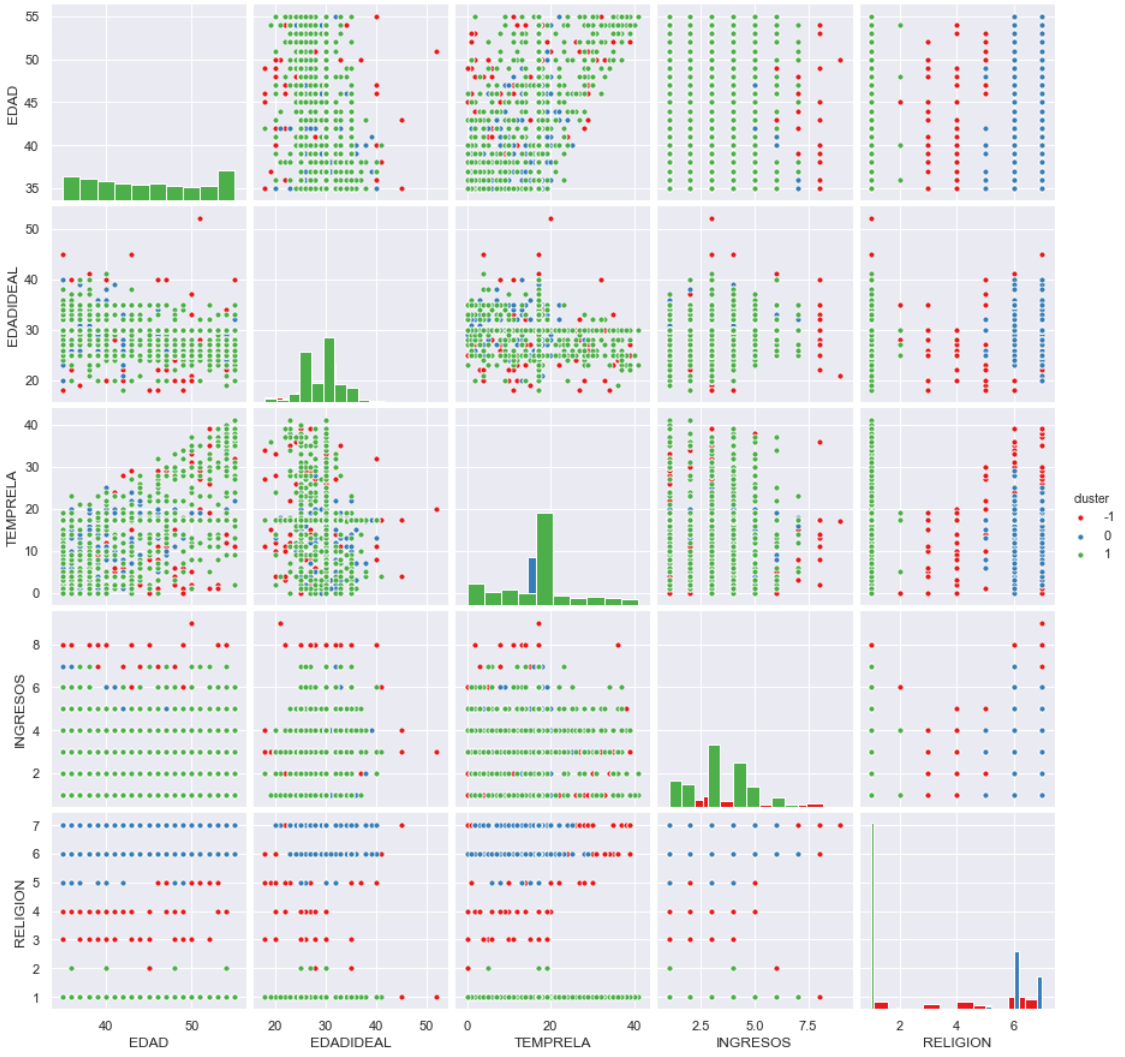
Nº clusters	Epsilon	Min_samples	T(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
3	0.2	10	0.02	0.30794	475.108	1: 654 (54.17 %) 0: 323 (28.23 %) -1: 167 (14.60 %)
2	0.3	10	0.03	0.14986	3.087	0: 1135 (99.21 %) -1: 9 (0.79 %)
2	0.2	50	0.02	0.16358	227.541	-1: 745 (65.12 %) 0: 399 (34.88 %)

La configuración con mejor Silhouette es la configuración $eps = 0.2$ y $min\_samples = 10$, el problema de esta configuración es que considera 167 datos como datos ruidosos. Las gráficas obtenidas con esta configuración son:



(g) Heatmap de DBSCAN del tercer caso de estudio.

Como se puede ver, vuelve a segmentar por la característica *RELIGION*.



(h) ScatterMatrix de DBSCAN del tercer caso de estudio.

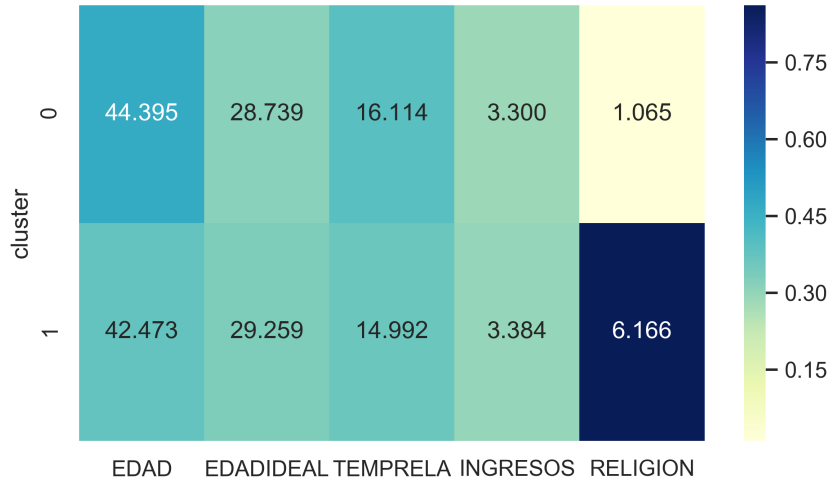
Los datos agrupados en el cluster -1, cluster de datos perdidos, son aquellos datos anómalos (*outliers*).

4.4. Birch

Se han probado las siguientes configuraciones:

Nº clusters	Umbral	Tiempo(s)	Silhouette	Calinski-Harabasz	Tamaño de cada cluster
2	0.2	0.05	0.43315	978.746	1: 683 (59.70 %) 0: 461 (40.30 %)
2	0.3	0.05	0.44032	998.578	1: 701 (61.28 %) 0: 443 (38.72 %)
2	0.4	0.05	0.44502	1012.037	0: 704(61.54 %) 1: 440 (38.46 %)
3	0.2	0.04	0.35630	645.876	0: 689 (59.70 %) 1: 304 (26.57 %) 2: 157 (13.72 %)
3	0.3	0.03	0.36881	677.936	1: 701 (61.28 %) 2: 267 (23.34 %) 0: 176 (15.38 %)

A la vista de los resultados, se elige la configuración de k=2 y umbral = 0.4.



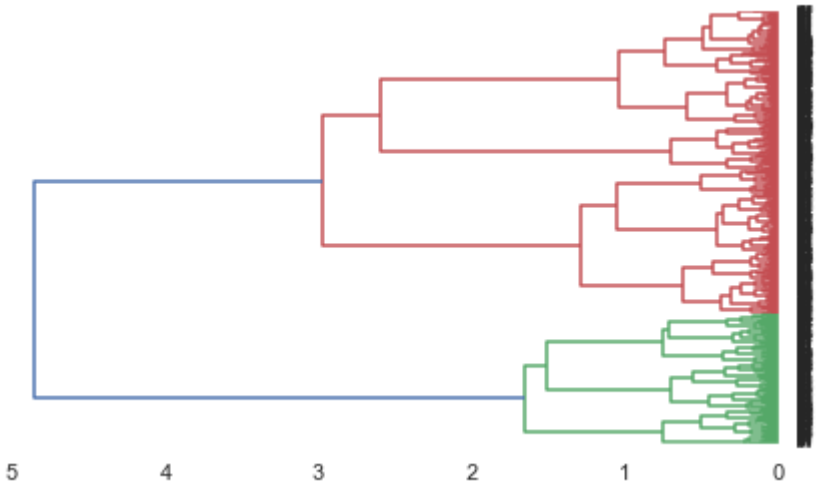
(i) Heatmap de Birch del tercer caso de estudio.

(j) ScatterMatrix de Birch del tercer caso de estudio.

Se logra una segmentación análoga a la obtenida por *Meanshift*.

4.5. Aglomerativo

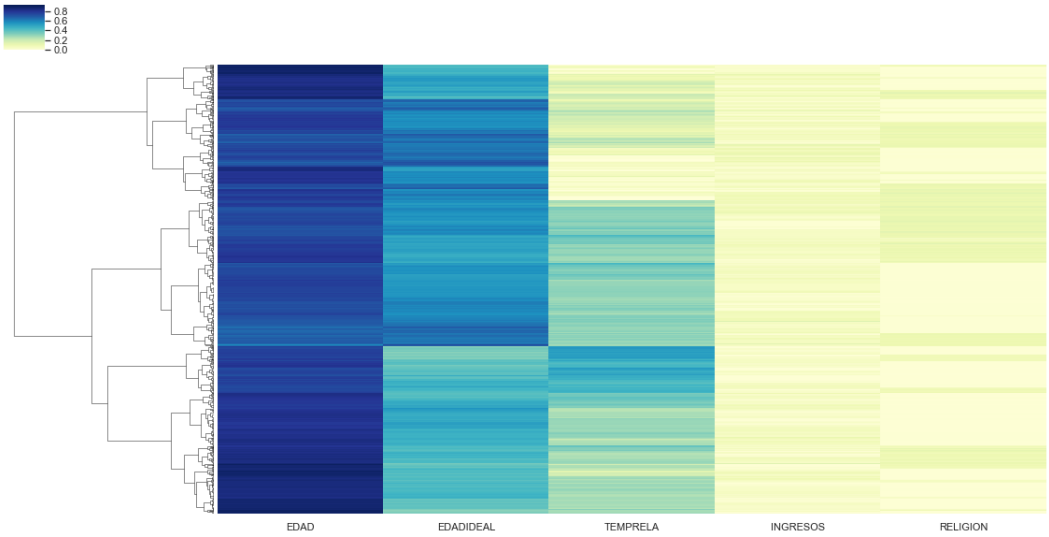
Para determinar el número de clusters de esta técnica se va a hacer uso del dendograma:



(k) Dendograma del tercer caso de estudio.

Aparantemente, viendo el dendograma se puede observar que con 3 clusters se logra distribuir bastante bien. Al asignar tres clusters, los resultados devueltos son los siguientes:

Nº clusters	Tiempo(s)	Tamaño de cada cluster
3	0.02	0: 372 (37.20 %) 2: 325 (32.50 %) 1: 303 (30.30 %)



(l) Clustermap del tercer caso de estudio.

Se puede observar como en un cluster la edad ideal es inferior que al resto de clusters. En el segundo cluster se tienen mujeres con relaciones de pareja más largas que en el tercer cluster. El resto de características, apenas se puede intuir la segmentación realizada.

4.6. Interpretación de la segmentación

A la vista de los resultados, se puede concluir varias afirmaciones:

- Las relaciones más longevas son de mujeres católicas, ateas o que han preferido no responder, mientras que las relaciones más cortas son de personas protestantes o musulmanas.
- La mayoría de mujeres coinciden en que la edad ideal para tener hijos es el intervalo [20,40] pero hay mujeres católicas que elevan esa edad ideal hasta los 50 años y, por otro lado, las mujeres protestantes acortan ese intervalo a los [25,35] años.
- Las mujeres más mayores, mujeres con 40 años de relación con una pareja, coinciden en que la mejor edad está en el intervalo de 20 a 30 años. Mientras que las mujeres, con menos años de relación, aumentan esa edad ideal hasta los 40 años. Esto puede ser debido a que las mujeres con menos años de relación aún quieren disfrutar de estar con su pareja sin hijos mientras

que las mujeres con más edad se han dado cuenta que los hijos proporcionan mucha felicidad y les hubiera gustado tenerlos antes.

5. Bibliografía

Referencias

[1] https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/, Consultado el 21 de Octubre.

[2] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html, Consultado el 21 de Noviembre.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html, Consultado el 10 de Noviembre.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html, Consultado el 10 de Noviembre.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html, Consultado el 10 de Noviembre.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html, Consultado el 10 de Noviembre.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html, Consultado el 10 de Noviembre.

[8] https://towardsdatascience.com/machine-learning-clustering-dbscan-determine-the-optimal-value-for Consultado el 23 de Noviembre.