Universidad Nacional Autónoma de México

Tarea 2

Analisis multivariado

Acosta Imandt Daniel

25 de noviembre de 2022

El reporte que viene a continuación se saco del análisis y los datos que se pueden encontrar en el siguiente enlace.

1.

Considere los datos correspondientes a los gastos que diferentes tipos de familias hacen en comida. Los tipos de familias son:
(MA) Trabajadores Manuales,
(EM) Empleados,
(CA) Jefes,
donde esta clasificación es de acuerdo a la actividad de quienes proveen el gasto para la casa, los datos provienen de familias en Francia. Por otra parte, en la clasificación también se considera el número de niños en la familia, de forma que MA2 denota a una familia donde el soporte económico lo proveen trabajadores manuales y la cual tiene 2 niños. Se tienen familias con 2, 3, 4 y 5 niños. Los tipos de alimentos son $X1$ = pan, $X2$ = vegetales, $X3$ = frutas, $X4$ = carne, $X5$ = pollo y aves, $X6$ = leche, $X7$ = vino. Haga un Análisis de Factores de estos datos. ¿ Cómo se relaciona este análisis con el NPCA que se llevó a cabo en el primer examen parcial ?

Solución

Primero vamos a ver las correlaciones entre los distintos productos que existen en la base de datos, con lo que se obtiene la siguiente matriz de correlaciones

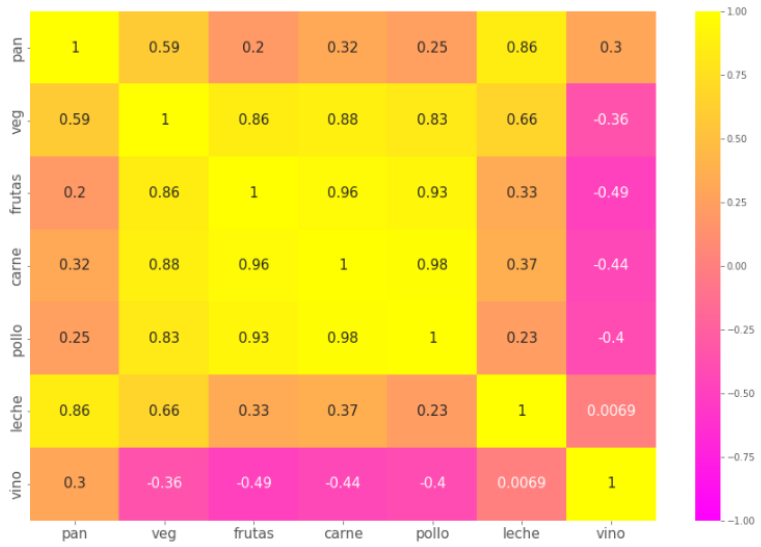

Figura 1: Matriz de correlaciones para food

En donde notamos algunos resultados bastante interesantes y es que parece ser que el vino no tiene una gran correlación ya sea positiva o negativa con ninguno de los otros productos, mientras que las frutas, la carne y el pollo están muy relacionadas, así mismo con los vegetales pero de una menor medida, por ultimo notamos que el pan solo parece estar bastante relacionado con la leche.
Por lo que si solo queremos tomar tres factores, podemos asociar el primero al pan,leche , mientras que el segundo factor a los vegetales,frutas,carne y pollo y el tercero a la variable del vino.
Así mismo notamos que cuando tenemos tres factores la varianza asociada es del 97 %, por lo que podemos reducir la dimensionalidad del problema manteniendo casi toda la información.
Una vez que se sacan las componentes principales vemos el valor que se le da a cada alimentos, con lo que se consigue la siguiente tabla.

	comida	componente 1	componente 2	componente 3
0	pan	0.169157	0.916757	-0.334536
1	veg	0.788544	0.533329	0.089904
2	frutas	0.927546	0.143073	0.274833
3	carne	0.972776	0.205934	0.096531
4	pollo	0.995686	0.076923	-0.027668
5	leche	0.164176	0.964148	0.166895
6	vino	-0.431351	0.186128	-0.581274

Figura 2: Matriz Q

Como se puede observar para la primera componente los vegetales,frutas y carne tiene un gran peso, mientras que para la segunda el pan y la leche son las que más valor tienen y por ultimo para la tercer componente el vino es el que más modifica lo esperado.

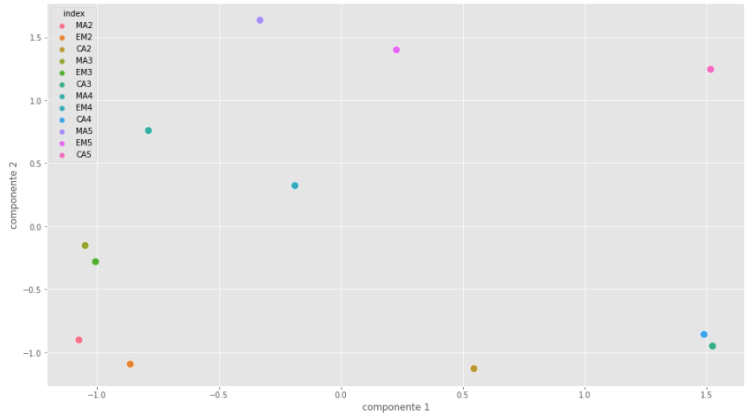

Figura 3: Componente 1 vs componente 2

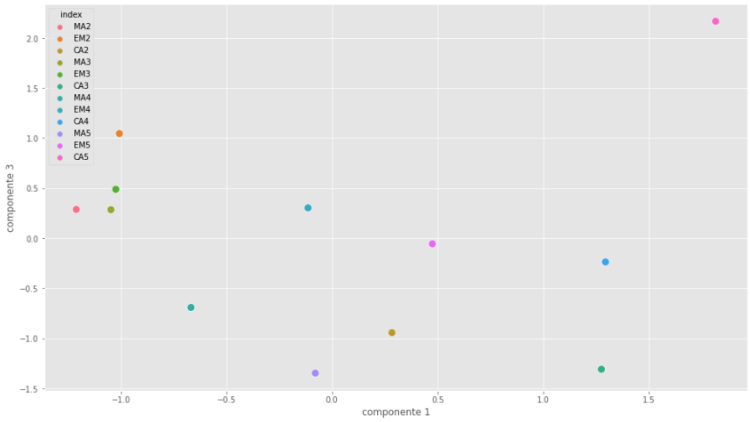

Figura 4: Componente 1 vs componente 3

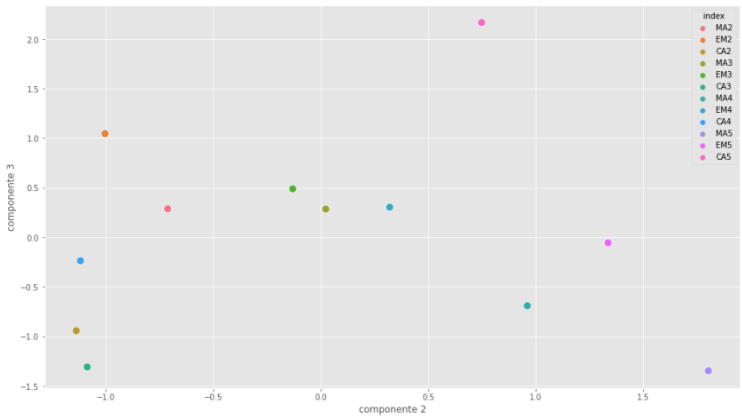

Figura 5: Componente 2 vs componente 3

A partir de todo esto podemos observar que las familias MA tienden a comprar mas comida de la componente 2, las familias CA se inclinan más por la componente 1.
Ahora recordando el análisis que se hizo en el examen, notamos que llegamos a resultados muy similares, ya que vimos que justo con tres componentes se llegaba a mas del 95 % de la varianza y que hay ciertas familias que se juntan dada su clasificación y el numero de hijos que tienen.

2.

Para los datos correspondientes a arrestos por diferentes crímenes en cada estado de los E.U.A. haga un análisis de conglomerados para cada método jerárquico mencionado en la Tabla 4.1 del libro de Brian Everitt Cluster Analysis (tabla entre las páginas 9 y 10 de las notas de clase). ¿ Son ciertas las afirmaciones sobre las propiedades de cada método jerárquico que se hacen en la quinta columna de esta tabla ? ¿ Qué método da un resultado más interpretable y porqué ?

Solución

Table 4.1 Standard agglomerative hierarchical clustering methods.

Method	Alternative name[a]	Usually used with:	Distance between clusters defined as:	Remarks
Single linkage Sneath (1957)	Nearest neighbour	Similarity or distance	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete linkage Sorensen (1948)	Furthest neighbour	Similarity or distance	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
(Group) Average linkage Sokal and Michener (1958)	UPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.
Centroid linkage Sokal and Michener (1958)	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals.
Weighted average linkage McQuitty (1966)	WPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven).
Median linkage Gower (1967)	WPGMC	Distance (requires raw data)	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals.
Ward's method Ward (1963)	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within clusters, after fusion, summed over all variables	Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers.

[a] U = unweighted; W = weighted; PG = pair group; A = average; C = centroid

HIERARCHICAL CLUSTERING 79

Figura 6: Tabla 4.1 del libro de Brian Everitt Cluster Analysis

A continuación se ve una muestra de como se ve la tabla de arrestos.

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
Iowa	2.2	56	57	11.3

Figura 7: Muestra de la tabla de los distintos arrestos

A continuación vamos a mostrar todos los análisis de conglomerados con los distintos métodos y vamos a ver las distintas propiedades que tienen.

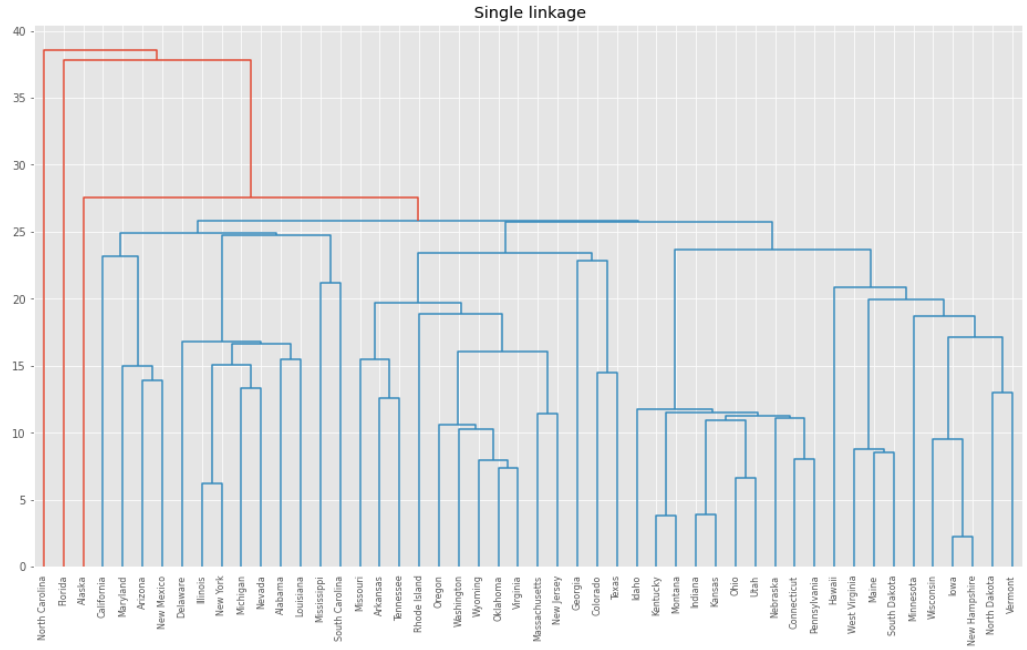

Figura 8: Single linkage

Según la tabla este método tiende a producir clusters desequilibrados y desordenados. (encadenamiento), especialmente en grandes conjuntos de datos. Por lo que no tiene en cuenta la estructura del conglomerado.

Ahora vamos a hacer un análisis de los datos obtenidos, en efecto notamos que los clusters se ven muy desequilibrados unos de otros y bastante desordenados, y no se tiene en cuenta como es la estructura de la tabla y del conglomerado, por lo que en efecto para el single linkage las afirmaciones de la tabla son verdaderas.

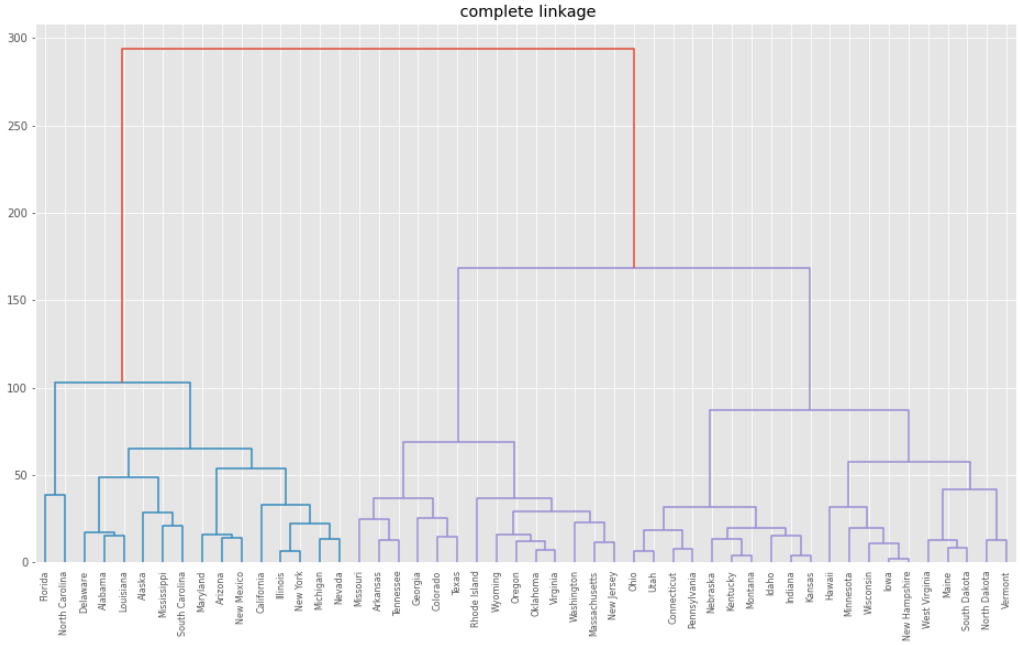

Figura 9: Complete linkage

Según la tabla este método tiende a encontrar grupos compactos con igual diámetro(distancia máxima entre objetos) No tiene en cuenta la estructura del clúster.

Notamos que las distnaica sy las formas de los clusters cambia bastante a cuando utilizabamos el metodo de "single", esto se debe a que en el pasado se agarraba la distancia minima, en cmabio ahora es la maxima, por lo que tiene sentido que sean bastante diferentes, por otro lado notamos que las distancias de los clusters son bastntre parecidas entre si, por lo que en efecto las afirmaciones de la tabla son ciertas.

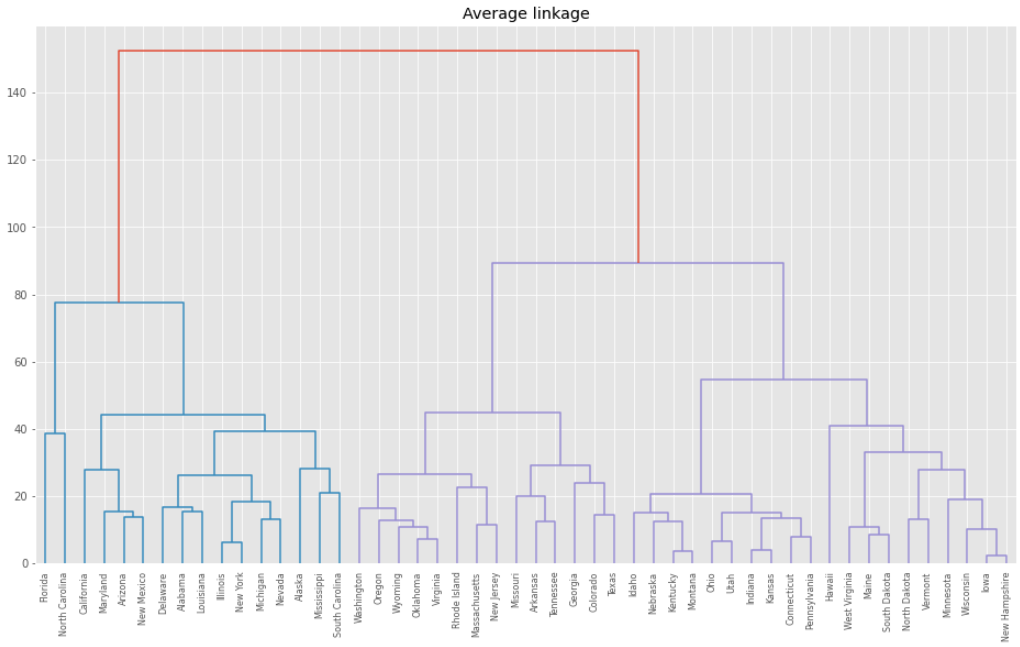

Figura 10: Average linkage

Según la tabla tiende a unirse a grupos con pequeñas variaciones.Es un intermedio entre 'single' y 'complete linkage'.Toma en cuenta la estructura del cluster y es relativamente robusto.

A partir de los resultados obtenidos notamos que se obtienen resultados muy parecidos a los del complete linkage, ya que los clusters se parecen bastante, lo que más cambia es la distancia la que los une, lo cual tiene sentido, ya que se ve el promedio de las distancias en comparación con el complete que ve la distancia máxima,así mismo notamos que si une a dos puntos cuando hay muy pequeñas variaciones entre ellos .

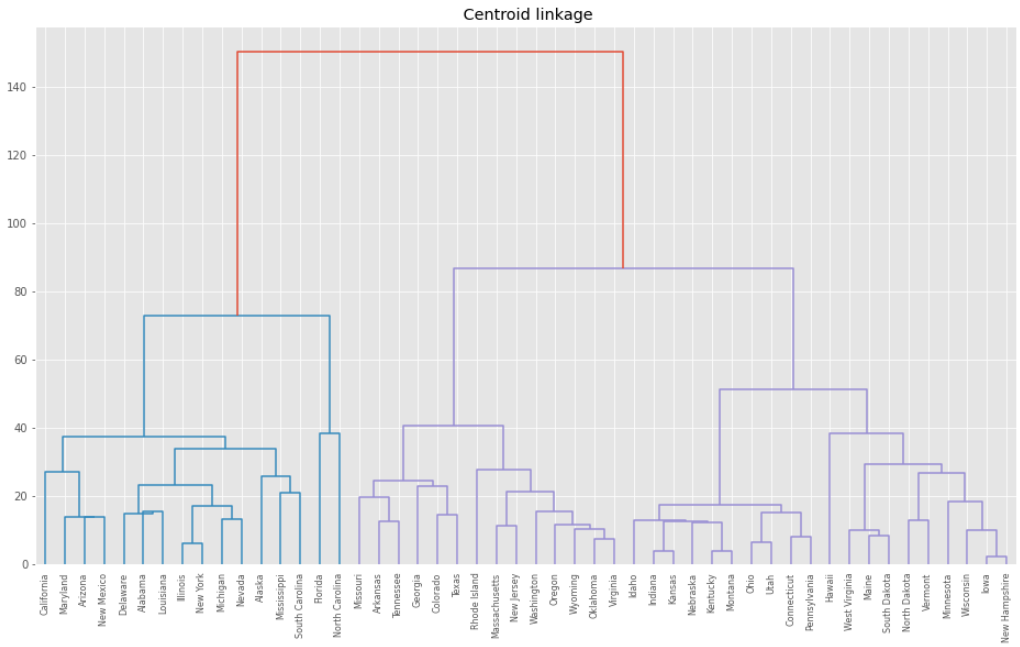

Figura 11: Centroid linkage

Según la tabla supone que los puntos se pueden representar en un espacio euclidiano.Cuanto más numerosos los dos grupos estos dominan el clúster fusionado.
Aquí logramos notar que en efecto hay un grupo más numeroso que otro y este es el que domina, por lo que si se cumplen las afirmaciones.

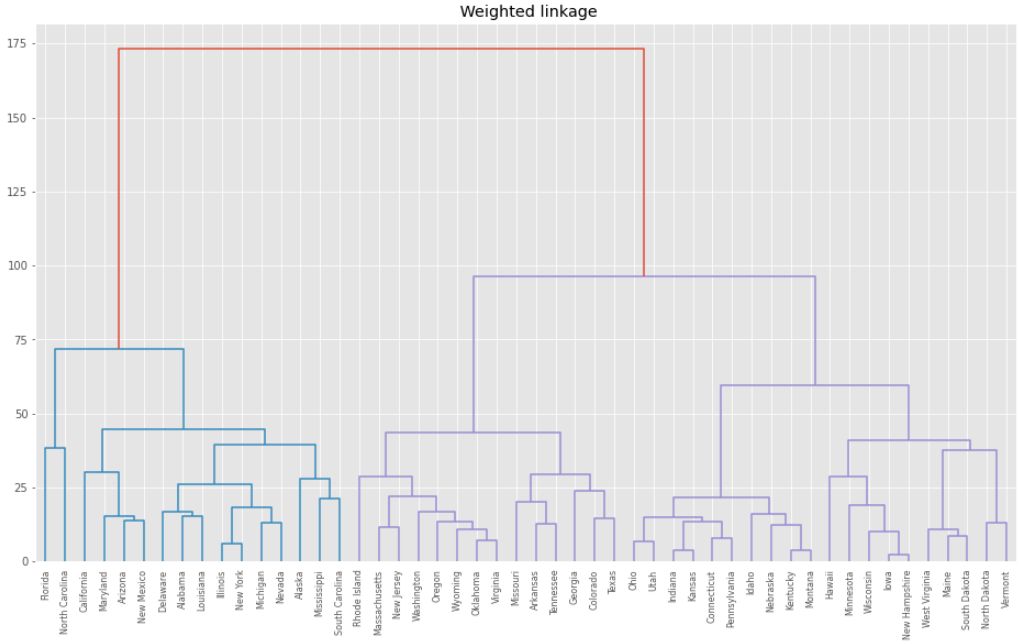

Figura 12: Weighted linkage

Según la tabla esta medida pondera más alto a los pequeños clústers que a los puntos en grandes conglomerados, esta medida nos sirve si es probable que los tamaños de los conglomerados no sean iguales.
Logramos observar que los clústers que menos puntos tienen en efecto son los que tienen una altura mayor y que el centroide no esta balanceado, por lo que la afirmación es cierta.

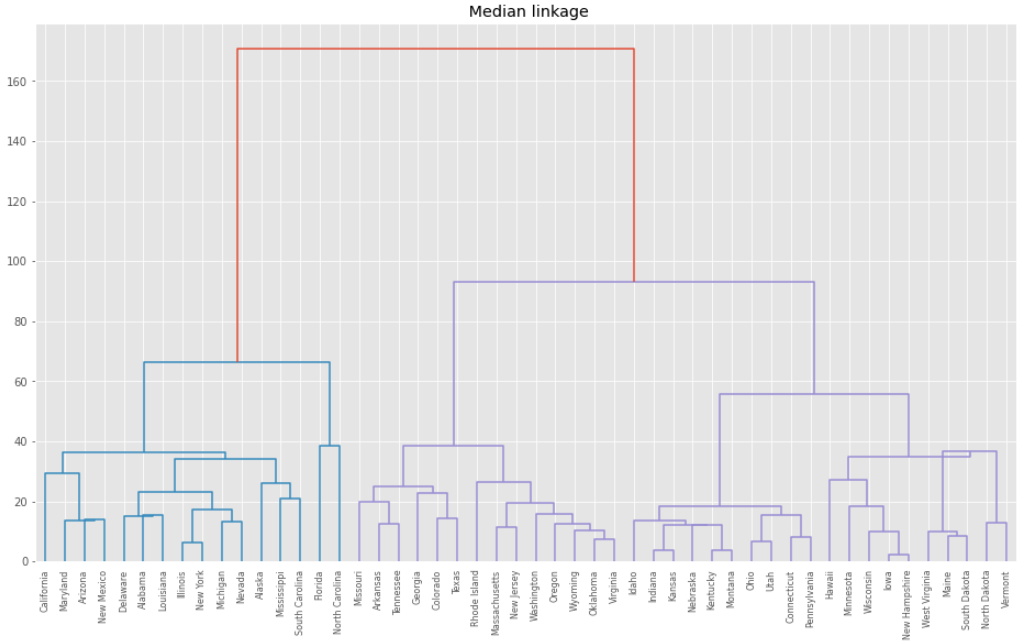

Figura 13: Median linkage

Según la tabla esta medida supone que los puntos se pueden representar en un espacio euclideano, el nuevo grupo es un intermedio entre la fusión de grupos.
Gracias a la figura logramos notar que en efecto los nuevos clústers se suelen generar a la mitad de los anteriores, por lo que las afirmaciones de la tabla son ciertas.

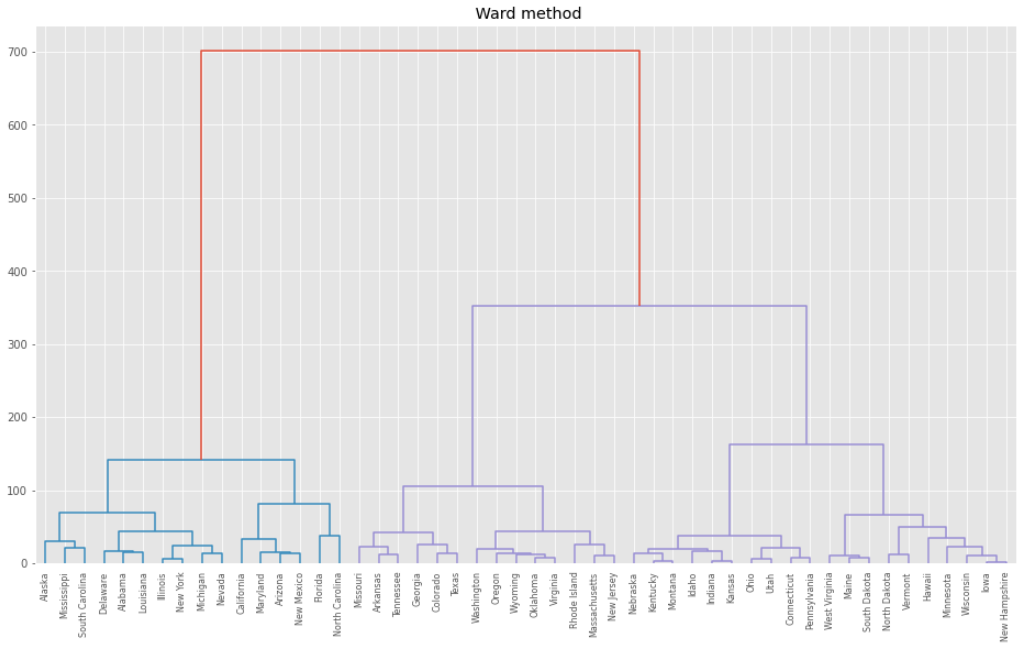

Figura 14: Ward method

Según la tabla supone que los puntos se pueden representar en un espacio euclidiano, tiende a encontrar grupos esféricos del mismo tamaño y es bastante sensible a valores atípicos.
Aquí logramos ver que los clústers suelen estar del mismo tamaño, y se logran buenos resultados, por lo que se podría decir que en efecto se cumplen todas las afirmaciones de la tabla.

Conclusiones ejercico 2

Después de hacer en análisis de conglomerados con los distintos métodos que se ven en la tabla 4.1, podemos decir que las afirmaciones de esta tabla son ciertas para los distintos métodos.
Ahora cada método sirve de forma diferente para cada conjunto de datos y para lo que se quiera llegar, por lo que cada método tiene sus ventajas y desventajas dado cierto conjunto de datos.
Pero por lo general logramos ver que el algoritmo de Ward tiene una forma más inteligente, ya que una dos grupos si el grupo resultante es lo más homogéneo posible y suele unirlos bastante bien, por lo que diría que es mi método de análisis de conglomerados favorito, pero este luego puede ser un poco más difícil de interpretar, que otros ya que es un poco más elaborado.
Por lo que diría que el método más interpretable es el de .ªverage linkageza que sigue una intuición que la gente suele utilizar bastante que es el de los promedios entre los puntos que están en cada clúster y así encuentra el punto medio de cada conjunto de puntos y a base de eso los decide unir, por lo que es fácil de entender y funciona por lo general mejor que el "single.º çompleteza que es es una mezcla de los dos y se puede explicar de una forma fácil sin tener que tener conocimientos extra de otros tópicos.