Tarea 2

Garduño Martínez Eduardo

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 n umero de ni nos en la familia, de forma que MA2 denota a una familia donde el soporte económico lo proveen trabajadores manuales y la cual tiene 2 ni nos. 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 ?

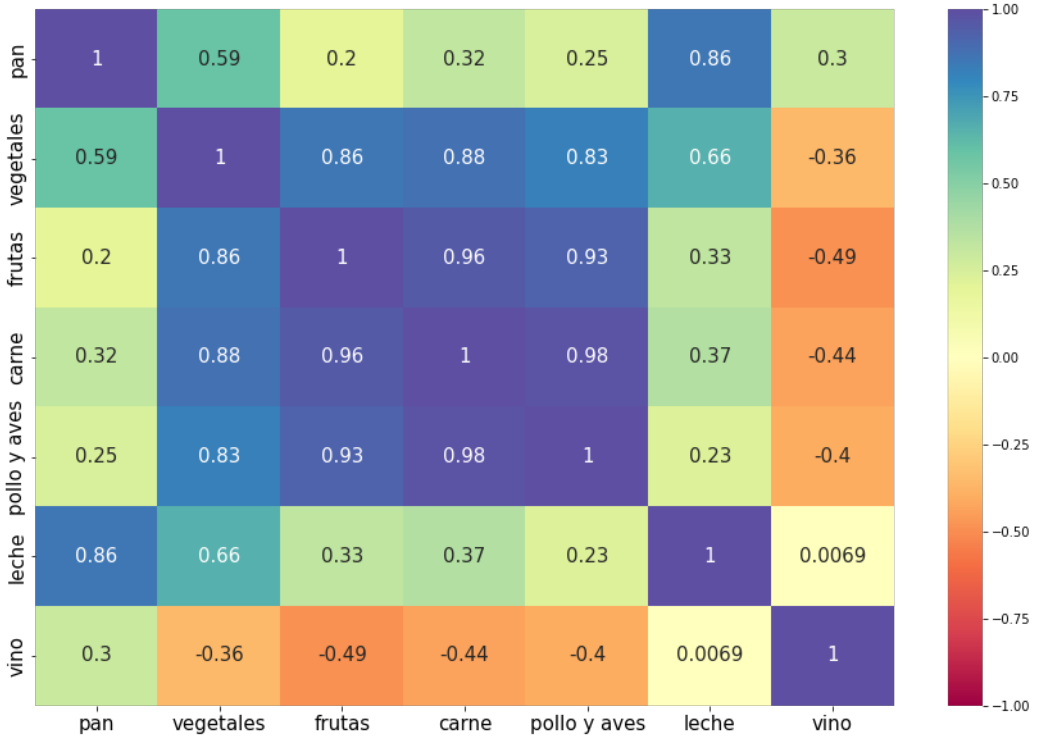

Figura 1: Correlación de las variables

Como se puede ver en el gráfico anterior podemos ver que hay múltiples variables correlacionadas como pueden ser el pan con la leche, mientras los vegetales tienen alta correlación con la fruta,carne y el pollo. Es así que en este problema si tiene sentido aplicar un análisis de componentes principales para poder reducir las dimensionalidad de nuestros datos, además viendo las correlaciones podemos sugerir en que se reduzca la dimensionalidad en 3 grupos que harían referencia al de (leche y pan); (vegetales, fruta, carne, pollo) y (vino), además de esto cuando usamos 2 componentes, entonces la varianza explicada seria del 88 % mientras que con 3 componentes tendríamos el 97 % por lo que en este análisis usaremos 3 componentes.

La matriz $\mathbb{Q}$ es la siguiente de los datos estandarizados:

pan	0.16915711927964786	0.9167574302464349	-0.3345364045707216
vegetales	0.7885439505652783	0.533328871966908	0.08990403935414862
frutas	0.9275456893765472	0.14307273212778363	0.2748327246876254
carne	0.9727755795998266	0.20593373885840136	0.09653117877599666
pollo y aves	0.9956864792115584	0.07692318357308713	-0.027668108203317287
leche	0.16417627622294886	0.9641479566405052	0.16689453701162624
vino	-0.43135051395474727	0.18612781036493467	-0.5812739024537299

Figura 2: Resultados de Q

Como podemos ver para la primer componente se puede relacionar con comidas más 'pesadas' o más 'fuertes', mientras que la segunda componente se pueden relacionar con el desayuno y por último la tercer componente se puede relacionar a la bebida.
Los resultados en el espacio de los factores se ven de la siguiente forma donde los puntos amarillos son los de clase CA, los azules EM y los morados a MA:

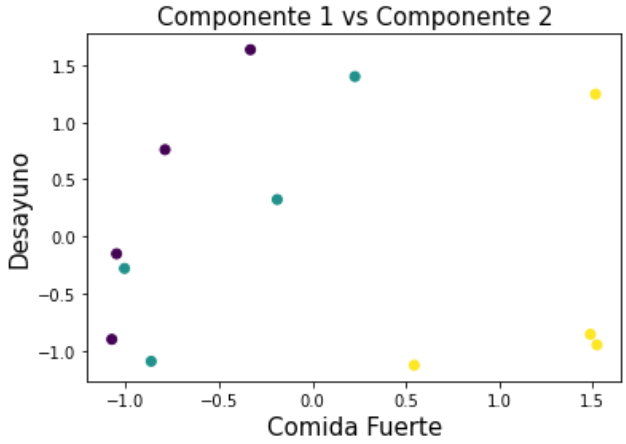

Figura 3:

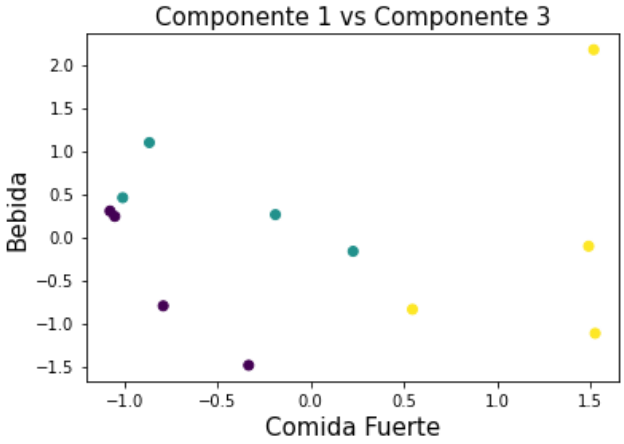

Figura 4:

Como podemos observar las familias CA poseen un mayor peso para las comidas 'fuertes' más que las otras, mientras que las familias MA tienden a comprar más comida relacionada con el desayuno que la familias EM, además de que las familias MA parece que tiende a gastar menos en Vino que la familias EM.

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?

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.

Figura 5: Tabla 1

Para poder discutir si son ciertas las afirmaciones de la tabla vamos a mostrar los resultados de los dendogramas. Como nota los datos se estandarizaron antes de realizar los dendogramas.

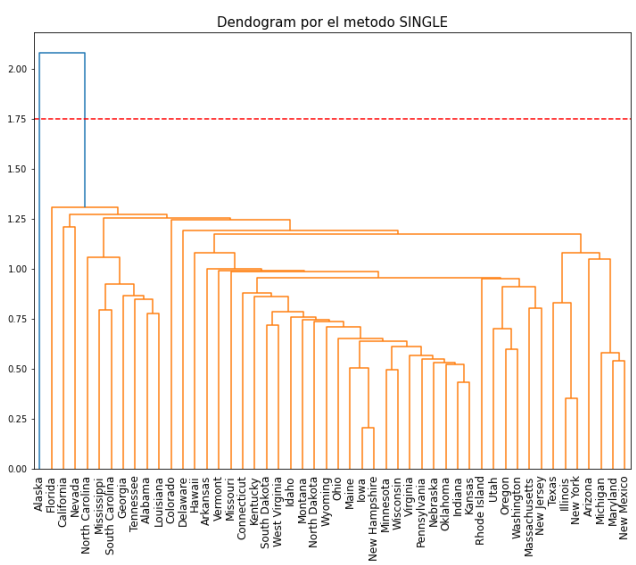

Figura 6:

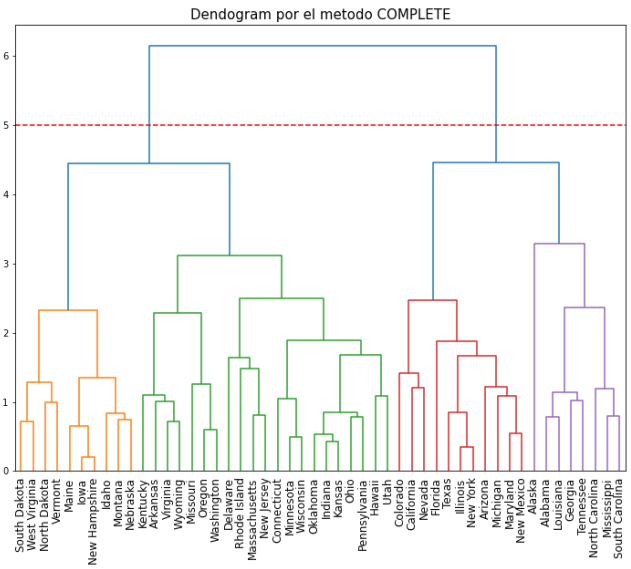

Figura 7:

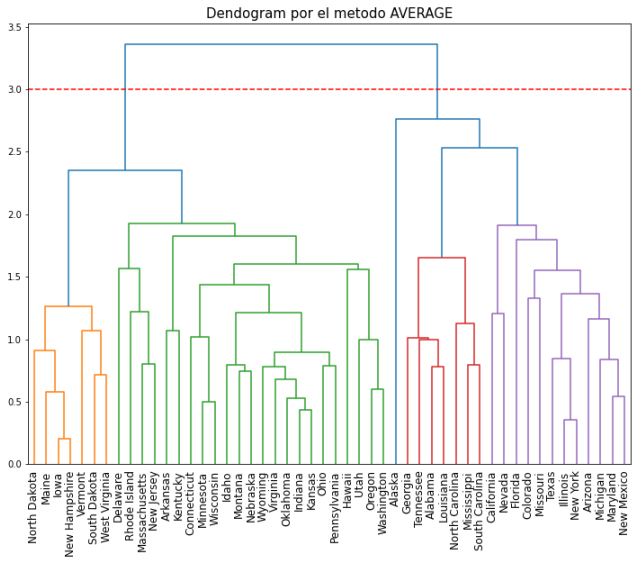

Figura 8:

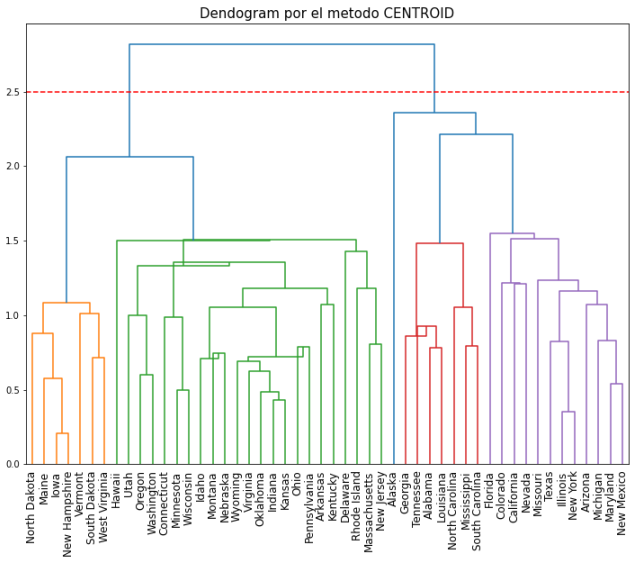

Figura 9:

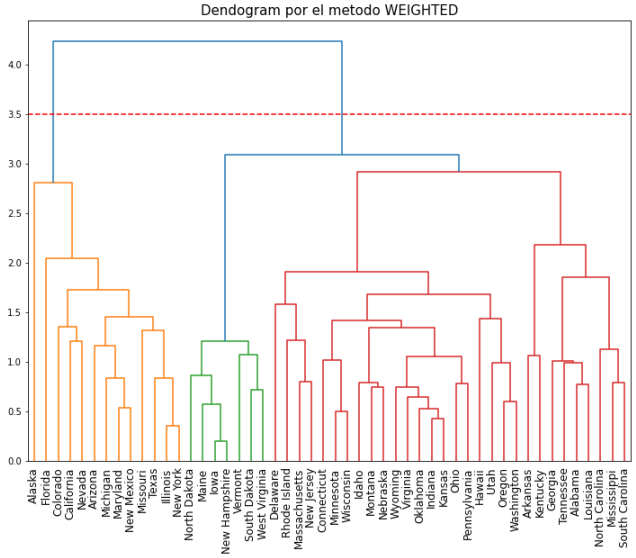

Figura 10:

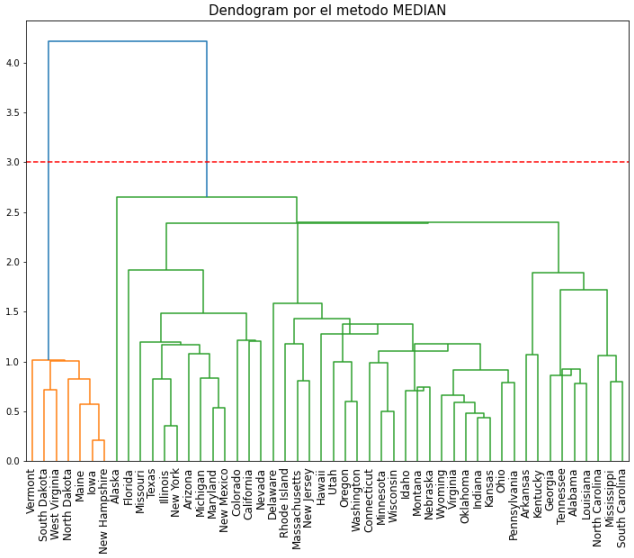

Figura 11:

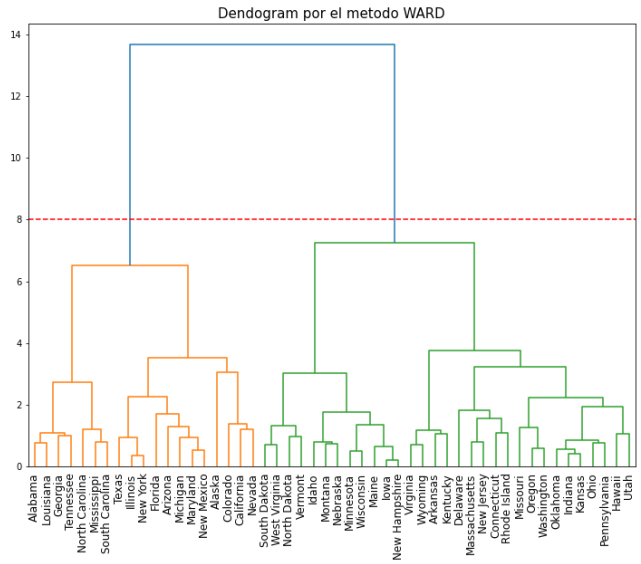

Figura 12:

Como podemos observar para el caso del método 'SINGLE' observamos que la aseveración hecha en la tabla es correcta ya que los clusters están desbalanceados ya que uno posee todos menos uno de los elementos mientras que el otro cluster solo posee 1. Para el caso del método 'COMPLETE' como podemos observar en la figura 7 los grupos amarillos, rojos y morados poseen una cantidad de elemento similares, mientras que el verde posee un poco más de elementos, además se ve que la máxima distancia de los elementos es similar si se toman 4 grupos, al igual que si se toman 2 grupos y se puede ver que la distancia no difieren de forma significativa por lo que a afirmación de la tabla parece ser correcta. Para el caso del método 'AVERAGE' parece producir clusters similares que el método complete aunque un poco más desbalanceados como puede ser el caso del conjunto verde y del morado que parecen tener más elementos por lo que si parece ser como algo intermedio entre single y complete. Para el caso de 'CENTROID' en este caso no hay mucho que decir lo resultados son muy similares al método average solo que algunas distancias parecen ser menores pero esto es debido a la naturaleza de como se hacen los enlaces. Para el caso de 'WEIGHTED' podemos ver que algunos clusters están desbalanceados ya que el cluster rojo posee más de la mitad de los estados, también podemos ver que las distancias máximas de dos de los clusters parece ser muy similares. Para el caso de 'MEDIAN' podemos ver que los grupos están muy desbalanceados ya que uno posee una gran cantidad de elementos mientras que el otro apenas posee 7 elementos. Para el caso de 'WARD' podemos ver que los clusters si parecen tener un tamaño similar por lo tanto esa afirmación de la tabla si parece ser correcta, sin embargo no podríamos a afirmar que los clusters son esféricos que ya no tenemos forma de visualizar los datos en 4D y verlo así.
Para responder ¿Qué método da un resultado más interpretable? Creo que depende de la forma original de los datos ya que si estos se agrupan en esferas entonces el método WARD ayuda a dar mejor interpretabilidad, en cambio si los clusters tienden a ser compactos, entonces el método COMPLETE será una mejor opción, mientras que si los clusters tienen poca varianza, entonces AVERAGE será mejor por lo tanto creo que esta pregunta se responde mejor conociendo los datos. Para nuestros datos creo que los mejores agrupamientos son AVERAGE, COMPLETE y CENTROID ya que esto parecen arrojar resultados similares, además de que nos muestran más clusters que los demás pero no demasiados por lo que esto nos podría ayudar a segmentar mejor a los grupos por ejemplo podriamos encontrar que un grupo es de los estados más 'seguros' o que un grupo se refiere a los estados con violencia intermedia,es decir, no son ni los más seguros ni los más violentos.