Examen práctico 1

Andrea Quintanilla

Marzo del 2021

Ejercicio 1

Usa herramientos que vimos en clase para entender mejor la estructura de este (sub)conjunto de datos.

Aunque nos faltan las herramientas de la tercera parte del curso, ¿qué harías para entender mejor las diferencias en las respuestas por género (usando lo que vimos)?

Nota. Los códigos de este ejercicio se adjuntan en el documento `examen_1_ejercicio_1.R`.

En `sog_agg_country.xlsx` los datos elegidos son porcentajes de respuestas binarias, de hombres de distintos países, a la pregunta:

Como resultado de la pandemia ocasionada por el coranovirus (COVID-19), ¿experimentaste algo de lo siguiente?

(health) Incapacidad de buscar asistencia médica.

(isolate) Aislamiento o seguimiento de cuarentena.

(job) Pérdida de trabajo.

(medical) Dificultad para obtener asistencia médica o artículos de higiene.

(migrate) Migrar a otra localidad.

(none) Ninguna de las anteriores.

(other) Otra.

(personal) Incapacidad de realizar rutinas usuales de cuidado personal.

(school) Actividades escolares canceladas o reducidas.

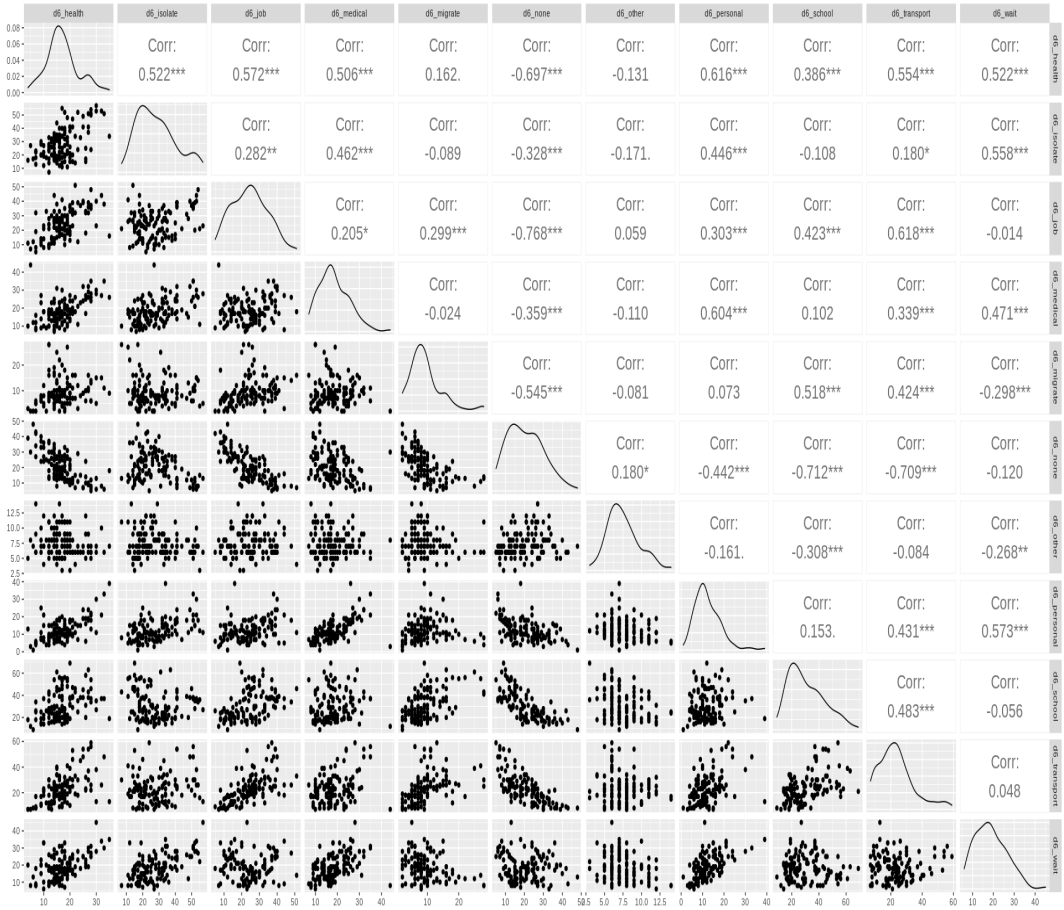
(transports) Inaccesabilidad a transporte público.

(wait) Esperas mayores para obtener asistencia médica.

Además de habernos restringido al subconjunto anterior, se omitieron las observaciones pertenecientes a regiones (y no a países específicos). Ya filtradas, los datos consistieron en 119 observaciones de 11 variables.

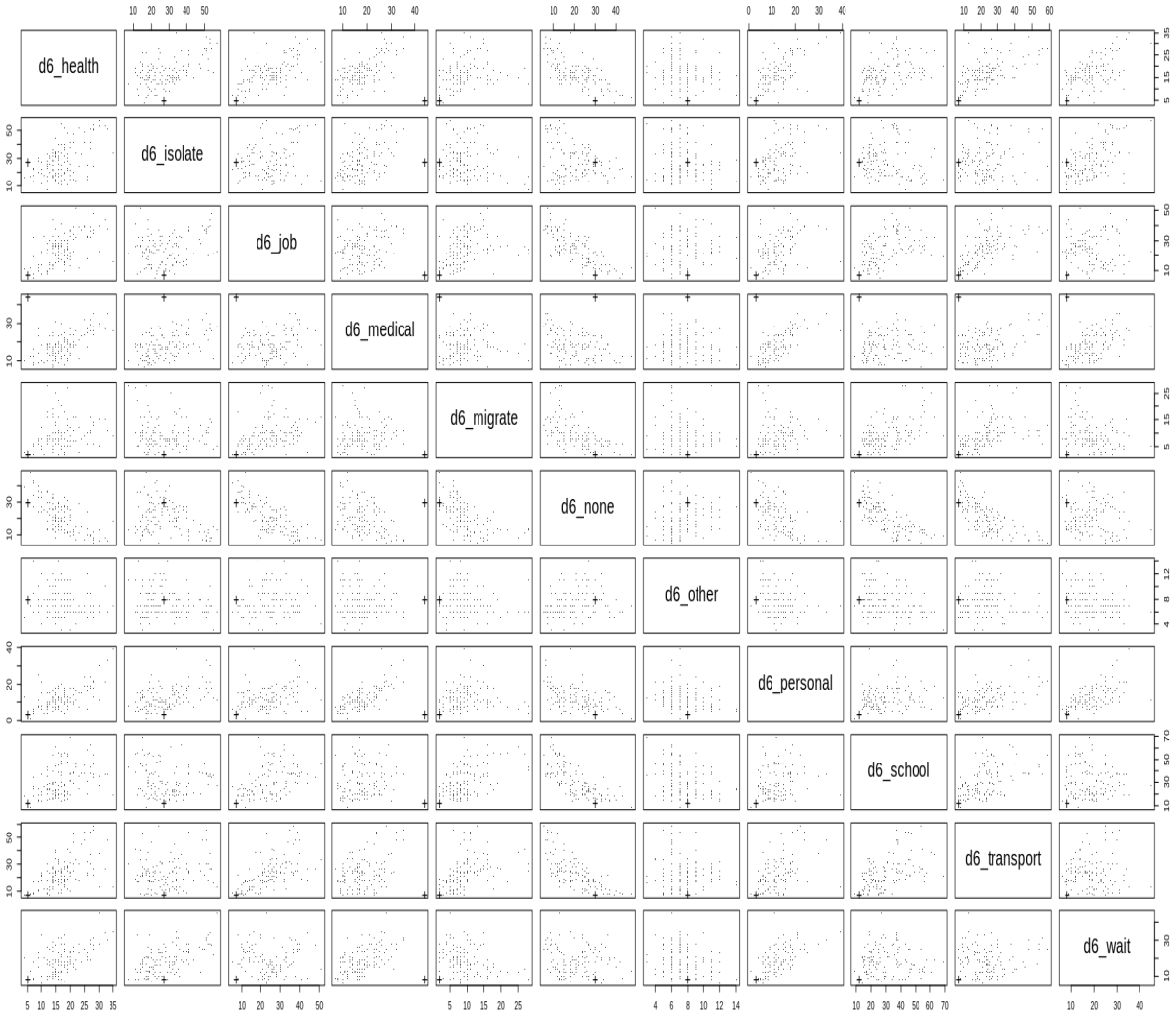
Para una primera exploración de los datos de utilizó la función `pairs()` y se obtuvieron las gráficas de las variables dos a dos. Se obeservaron algunas relaciones sugieren linealidad, por ejemplo, *health* con todas las variables, salvo *other*. Mientras que otras se alejan de ella, como la variable *other*. En esta misma gráfica las marginales parecen ser unimodales y las correlaciones varían aproximadamente entre 0 y .6.

Figura 1: Pairs-plot.



También se buscaron datos atípicos. Tanto en las gráficas de las variables *medical*, *personal* y *wait*, se observa un punto que está relativamete separado. Se identificaron como los países de Japón, Lituania y Puerto Rico, respectivamente. En las siguiente gráfica se repite el pairs-plot pero se señala a Japón con una cruz para identifdicar si es que su comportamiento es atípico en todas las variables. Gráficas análogas se hicieron para Lituania y Puerto Rico. En el caso de Puerto Rico se encontró sólo se separa fuertemente en la variable *wait*. Mientras que en los otros dos casos parece ser que comportamiento es atípico sólo en dos variables; *health* y *personal* para Lituania y *health* y *medical* en el caso de Japón. Debido a esto se decidió en un primer análisis continuar con los 3 países.

Figura 2: Pairs-plot y outliers. Japón con una cruz.



Al aplicar PCA se encontró que con las primeras dos componentes se explicaba alrededor del 62 % de la varianza. Para visualizar las contribuciones de cada estación a esas componentes graficamos $(i, p1_i)$ y $(i, p2_i)$, con $p1$ y $p2$ la primera y segunda componente, respectivamente.

Figura 3: Curva de pesos de la primera componente principal

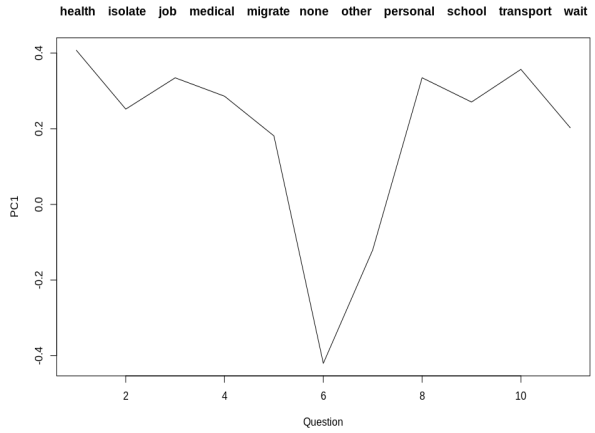
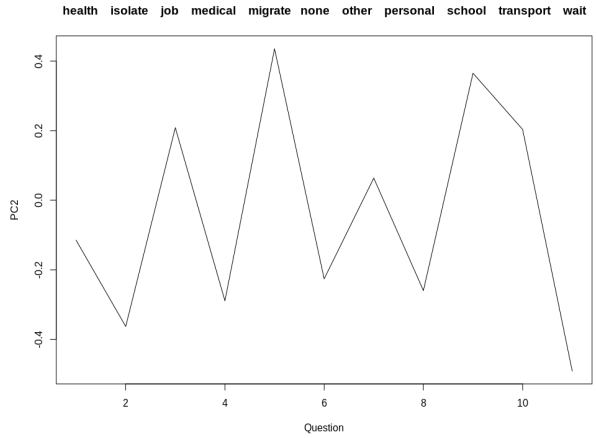


Figura 4: Curva de pesos de la segunda componente principal



Estas gráficas nos indican que la primera componente principal es un contraste entre la variable none y algo similar al promedio de las demás variables (health, isolate, job, medical, migrate, personal, school, transport y wait), salvo la variable other, pues ésta tiene un peso cercano a cero. Esto es razonable pues las personas que resultan no tener ningún problema médico, económico mayor, se dsitnguen de los que sí sulen tener alguna por que un rezago (económico, educativo, médico) suele venir acompañado por otros.

Sobre la segunda componente, los loadings indican un contraste entre el promedio de las variables wait, isolate, medical y personal por un lado, con el promedio de las variables school, migrate, transport y job. El último grupo tiene que ver en mayor medida con pérdidas económicas (perder trabajo, suspensión de actividades escolares y transporte, tener que migrar). Mientras que el primero tiene que ver mayormente con asuntos médicos-higiénicos (esperar por asistencia médica, incapacidad de buscar asistencia médica, incapacidad de rutinas normales de aseo), más la variable de aislamiento.

Las variables other y health tienen un peso cercano al cero. Esto último se explica porque son las que están mayormente representadas en la primera componente.

A continuación se muestra la proyección de los datos originales en las primeras dos componentes. Para revisar si se había capturado cierta estructura, se colorearon las regiones. Notamos que efecitvamente, dichas variables tienen una relación fuerte con la región y que las componentes principales logran resumir parte de estas interacciones. Junto con el biplot que se muestra enseguida podemos interpretar más fácilmente estas proyecciones. Las regiones de África Sub-Sahariana, de Latinoamérica y el Caribe, presentan mayores respuesta afirmativas a falta de transporte, pérdida de trabajo. La región de África-Subsahariana es la que tendió a migrar más y al cierre de escuelas. Latinoamérica y el Caribe fueron las que se vieron más afectadas en el sentido médico-higiénico y por situaciones de aislamiento/cuarentena. Mientras que en el extremo contrario, los que están positivamente más relacionados con la variable None, los que no se vieron afectados, están las regiones de Europa, Asia Central y Norte América, junto con algunos países de Asia del Este, como Japón, que es potencia, Australia y Taiwán.

Figura 5: Gráfica de proyección en PC1 y PC2. Coloreado por región.

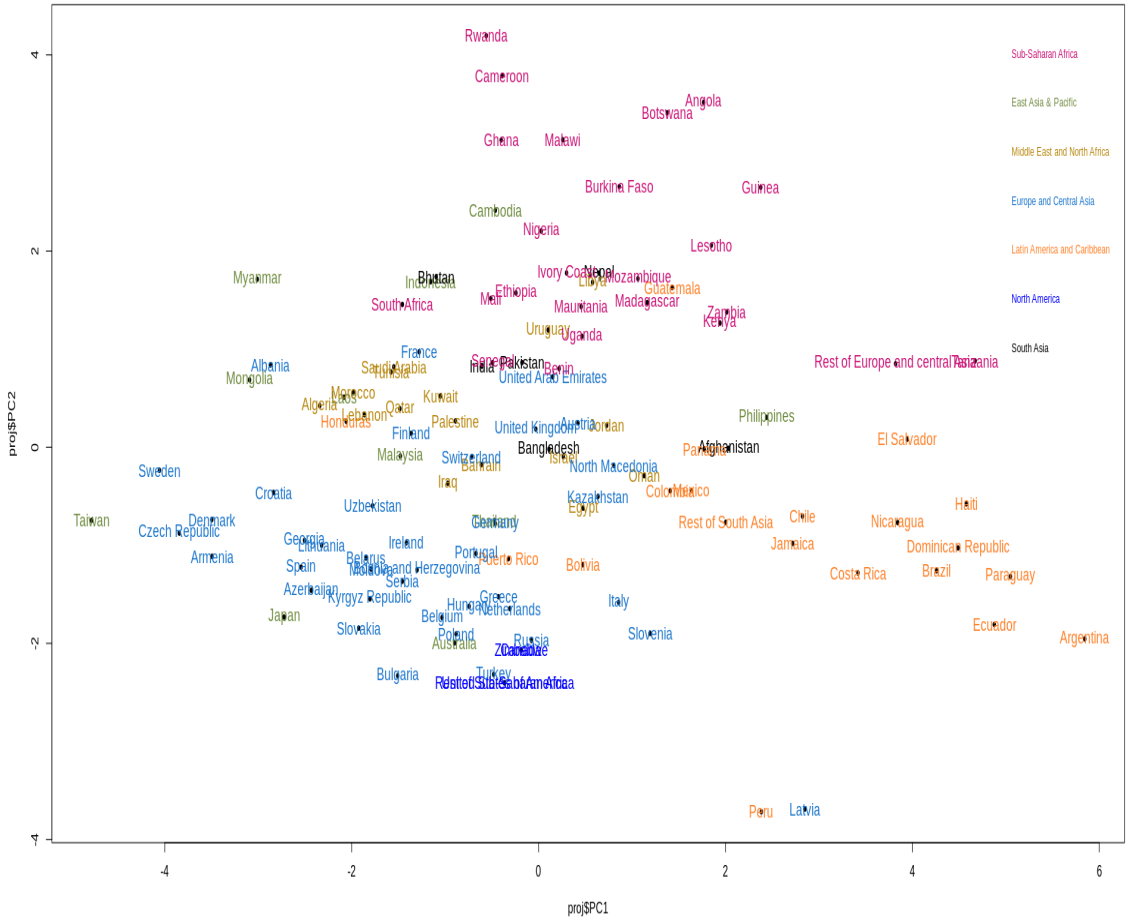
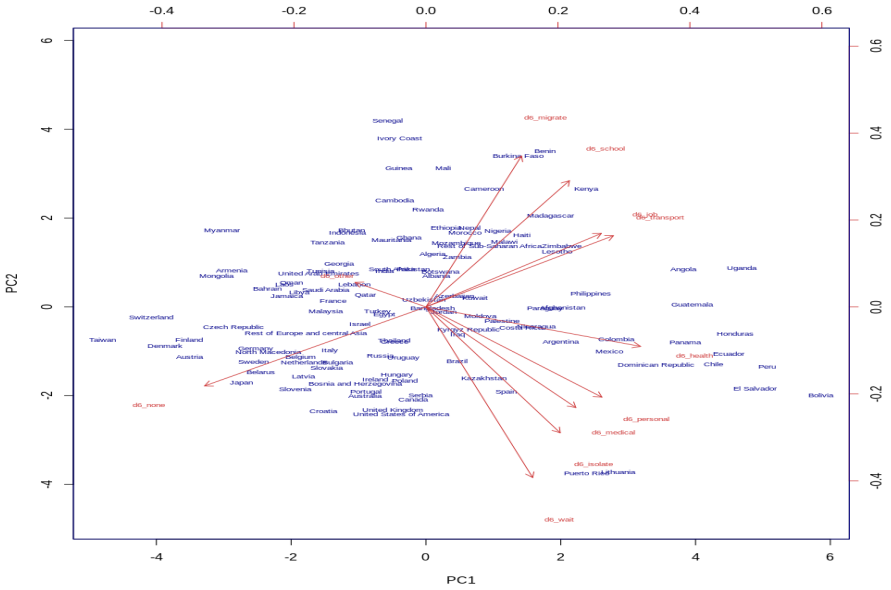


Figura 6: Biplot



Finalmente, para entender mejor las diferencias en cuanto a género, podríamos repetir el mismo análisis pero considerando sólo a mujeres y ver si son similares los loadings. También podríamos incorporar todos los datos y revisar con técnicas de agrupamiento si éstas sugieren dos grupos.

Ejercicio 2

Considera los datos del archivo `food.txt` se trata de 961 alimentos, su peso y sus componentes nutriciales. La finalidad es entender si hay grupos (clusters) presentes en estos datos. Para eso, busca visualizaciones informativas y usa técnicas de clustering (hint: como el peso varia, es mejor normalizar dividiendo por el peso; estos son los valores en las últimas 6 columnas).

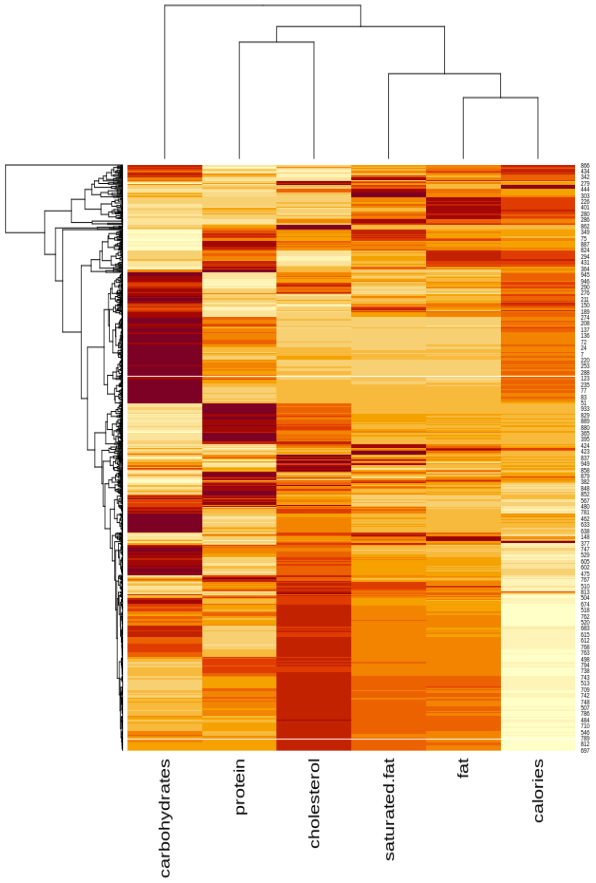
Nota. Los códigos de esta sección se adjuntan en el documento `examen_1_ejercicio_2.R`.

Como se menciona en la sugerencia, se utilizaron las últimas 6 columnas de datos. Así, el conjunto de datos trabajados, consiste en 961 observaciones correspondientes distintos alimentos y en sus valores nutricionales en las variables: grasa, calorías, carbohidratos, proteína, colesterol, y grasas saturadas.

Se comenzó construyendo un mapa de calor. Éste ocupa agrupamiento jerárquico para ordenar tanto las columnas como las observaciones en los datos. Colores obscuros corresponden a valores relativamente altos (rojos) y colores claros (amarillos) a valores bajos. Así, observaciones cercanas estarán de manera contigua. Por ejemplo, si nos restringimos a la columna de carbohidratos se aprecian tres grupos obscuros, tres claros y algunas zonas intermedias (naranjas). En la variable de colesterol podemos notar dos grandes grupos, uno obscuro en la parte inferior y uno claro en la superios. Igualmente en calorías parecieran existir dos zonas, una clara y una obscura. La variable de proteínas está mayormente fragmentada. Mientras que en las dos columnas de grasa existen dos grandes regiones claras e intermedias y varios fragmentos.

Pareciera que la única agrupación de los alimentos que se mantiene a través de varias variables es en la de colesterol, grasas, grasas saturadas y calorías. Pues estas cuatros categorías tienen dos zonas

Figura 7: Heatmap y dendogramas.



amplias claras y obscuras. Observamos que esas esas cuatro variables tienen el mismo de orden de claro a obscuro, salvo la de calorías, así para fines de visualización, se invirtieron los valores de calorías de x a $-x$. También podemos observar que la última rama del dendograma correspondiente a las observaciones es un dato aislado, por ello, éste se removió. El heatmap resultante tras esas modificaciones se muestran a continuación.

Éste ahora sugiere más fuertemente al menos dos grandes zonas, una en la zona inferior y otra en la superior. Pensando en ellos, se eligió un representante de cada una de esas zonas (a la obsevación 90 y a la 471) y se hicieron suficientes cortes como para que quedaran en distintos clusters: 13 clusters (13 cortes para separar la zona clara de calorías de la zona obscura). Los resultados se muestran en un dendograma radial (como un árbol filogenético).

Figura 8: Heatmap y dendogramas modificado.

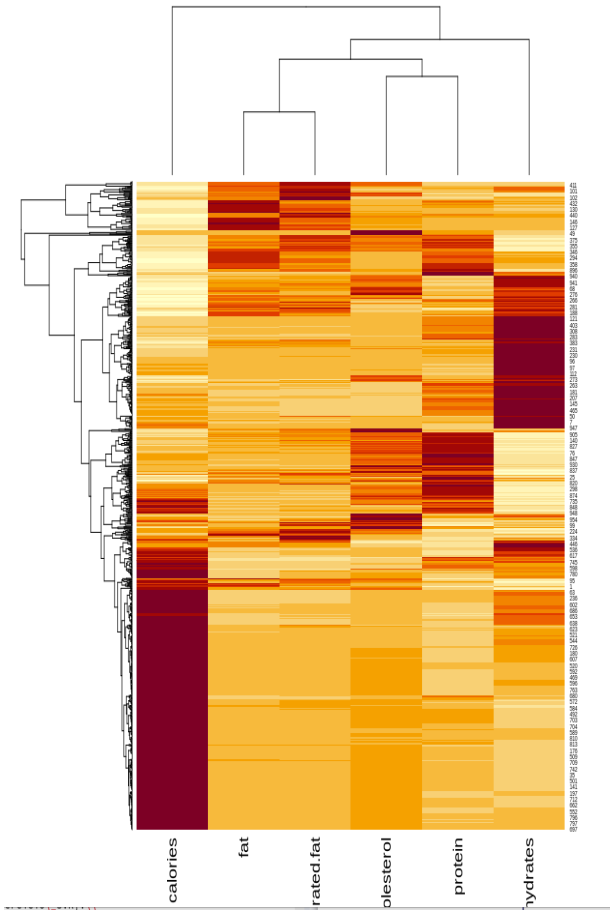
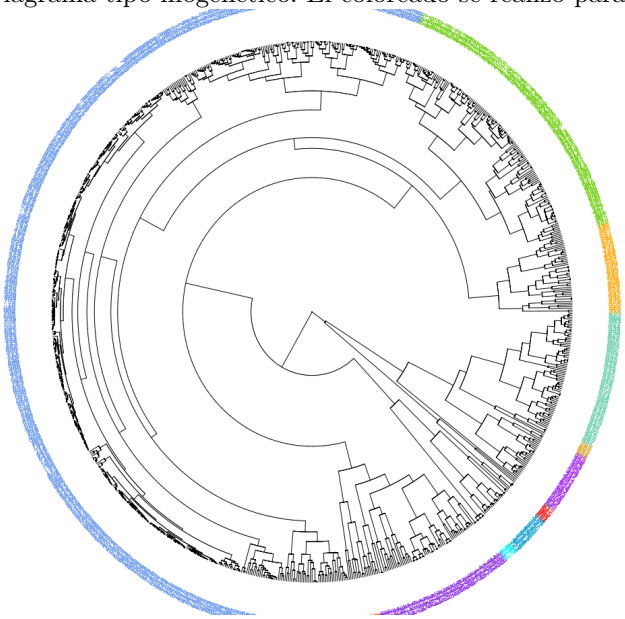
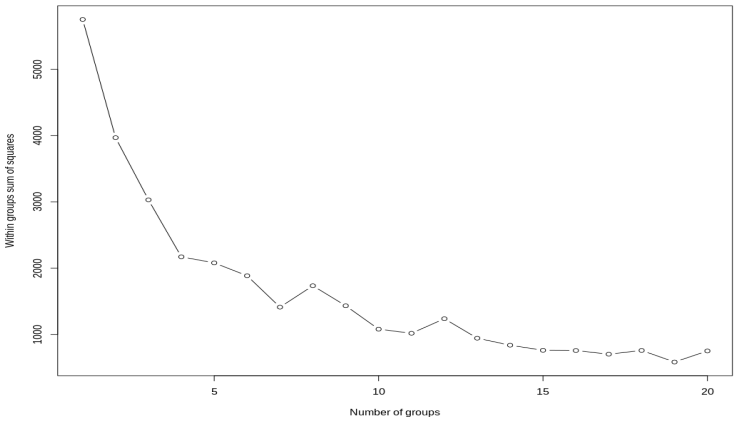


Figura 9: Diagrama tipo filogénetico. El coloreado se realizó para 13 clusters.



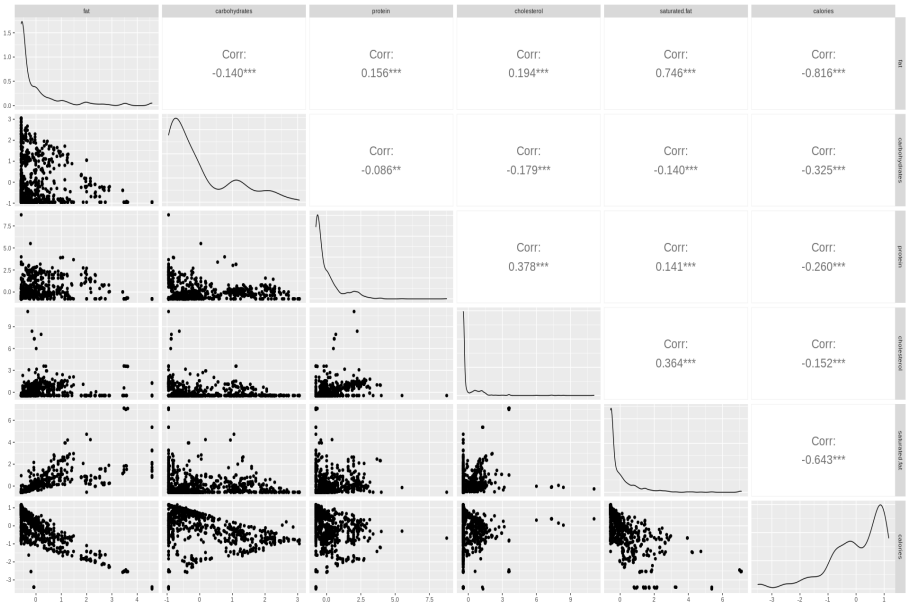
Después, se aplicó k-medias. En la gráfica siguiente, donde se ilustra el decrecimiento de la varianza interna de grupos media vs. k para clusters entre 1 y 20, se muestran resultados congruentes con la cantidad de cluster obtenidos anteriormente, la suma después de 11-12 clusters se estabiliza.

Figura 10: WGSS vs. K, para K-medias.



Por lo anterior se decició continuar con 12 clusters. Para visualizar las agrupaciones se proyectaron los datos utlizando PCA y se colorearon sengún su grupo. En la siguiente gráfica se observan los pares de variables; las relaciones no son lineales, en especial en algunas de las gráficas correspondientes a colesterol y proteínas, pero en otras gráficas se acercan a ella y casi todas las distribuciones son unimodales.

Figura 11: Pairsplot, distribuciones y correlación.



Al realizar PCA la varianza explicada por las primeras dos componentes fue del 67 %. Para interpretar dichas componentes mostramos de nuevo las gráficas de variables vs. loadings. La primera componente parece ser un contraste entre calorías y grasas (saturadas y no saturadas). Y la segunda, otro contraste entre carbohidratos por un lado y un promedio entre proteínas, calorías y colesterol por el otro.

Figura 12: Curva de pesos de la primera componente principal

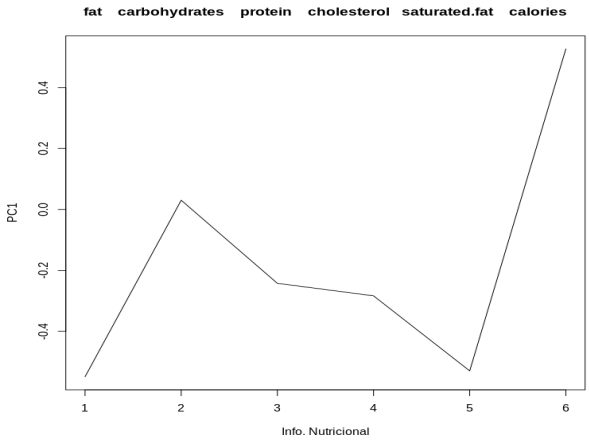
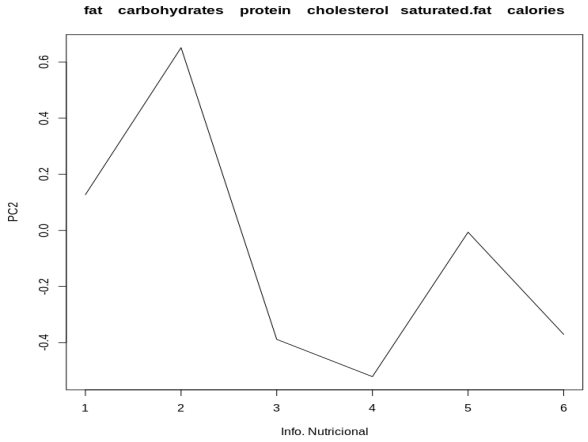
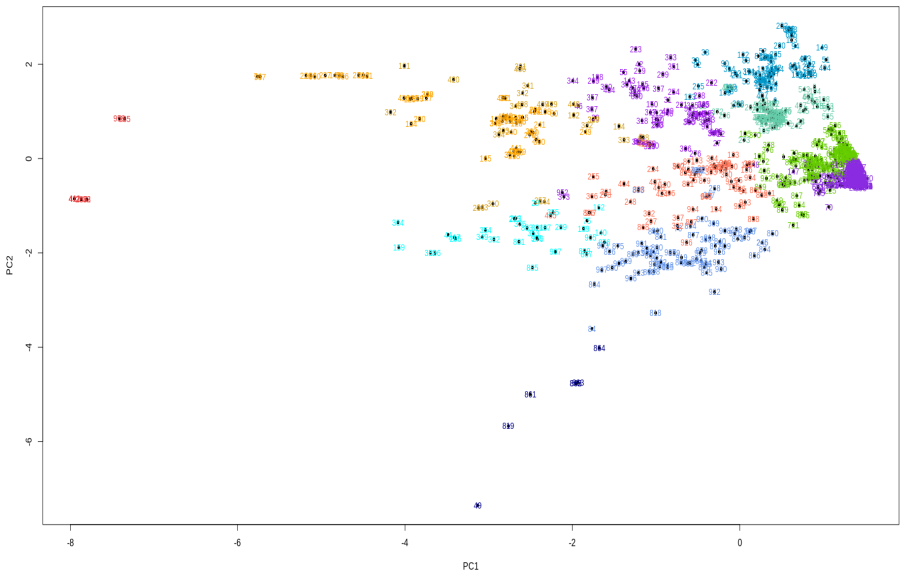


Figura 13: Curva de pesos de la segunda componente principal



Finalmente, coloreamos las proyecciones de los datos por cluster. En esta visualización se observa bastante estructura pues todos los colores están agrupados, salvo el morado que se fragento en dos pedazos. Dicha fragmentación se podría deber a que la agrupación no fue adecuada y se necesitaba un número mayor de clusters o, probablemente (porque no se cumplen todas las hipótesis de PCA) a que la proyección no es adecuada. Hubiera sido interesante tener los nombres de los alimentos para elegir algunos representantes por grupo e interpretar más los resultados.

Figura 14: Visualización de agrupamiento con k-medias, para $k = 12$, en las proyecciones de las primeras 2 componentes principales.



Ejercicio 3

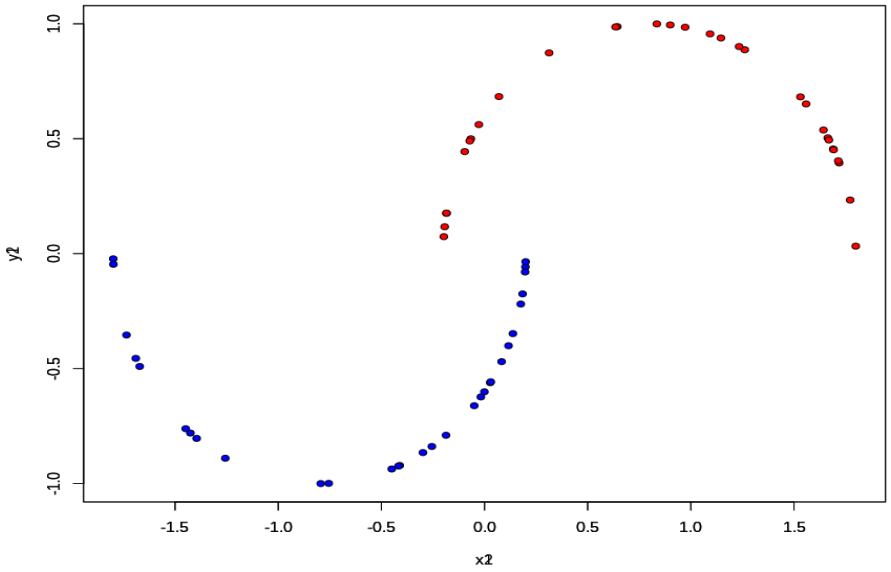
Implementa agrupamiento espectral no normalizado. Escribe un pequeño reporte didáctico que muestra y discute con ejemplos el desempeño de este algoritmo.

Nota. Los códigos de esta sección se adjuntan en el documento `examen_1_ejercicio_3.R`.

Para esta implementación se utlizó como matriz de similitud la de kernel gaussiana. Como se observará en los figuras que siguen, resultó ilustrativo elegir dicha medida de similitud pues permitió explorar la sensibilidad del algoritmo a los pesos asignados, aprovechando que el parámetro δ se puede variar para simular más o menos cercanía entre los datos.

Como ejemplo se tomaron las semilunas utilizadas para explorar TSNE en una tarea pasada. Éstas se contruyeron eligiendo puntos en dos parametrizaciones de semicírculos uniformemente. Cada semicírculo contiene 30 puntos. Se eligió este conjunto pues una ventaja de este algoritmo es que no pide una estructura a los datos, en contraste, por ejemplo, con PCA, que supone una relación lineal y falla mientras se alejan de ella.

Figura 15: Semilunas uniformes. Datos para ejemplificar el algoritmo. 60 observaciones



En el programa implementado se siguieron los siguientes pasos:

- Cálculo de la matriz de similitud con kernel gaussiano de parámetro δ.
- Cálculo de matriz Laplaciana. Y visualización con levelplot.
- Extracción de los primeros dos eigenvectores menores con la función `eigen`.
- Gráfica de *Within groups sum of squares* (WGSS) vs. K del agrupamiento K-medias, sobre las filas de los eigenvectores extraídos en el conjunto anterior. Y agrupamiento K-medias en dichos datos.
- Visualización del agrupamiento en el conjunto de datos original.

Se experimentó con valores de δ entre 0.1 y 15, que resultaron en un diversos comportamientos. En todos los casos los grupos se identificaron correctamente en los extremos más alejados de las semilunas, pero en los datos centrales existía mayor confusión para valores muy cercanos a los extremos (0.1 o al 15). Para valores alrededor del 3 la clasificación era más cercana a la deseada. Esto se muestra enseguida:

Figura 16: Visualización del agrupamiento en los primeros 2 eigenvectores menores en el conjunto de datos original. Para $\delta = 0.2$

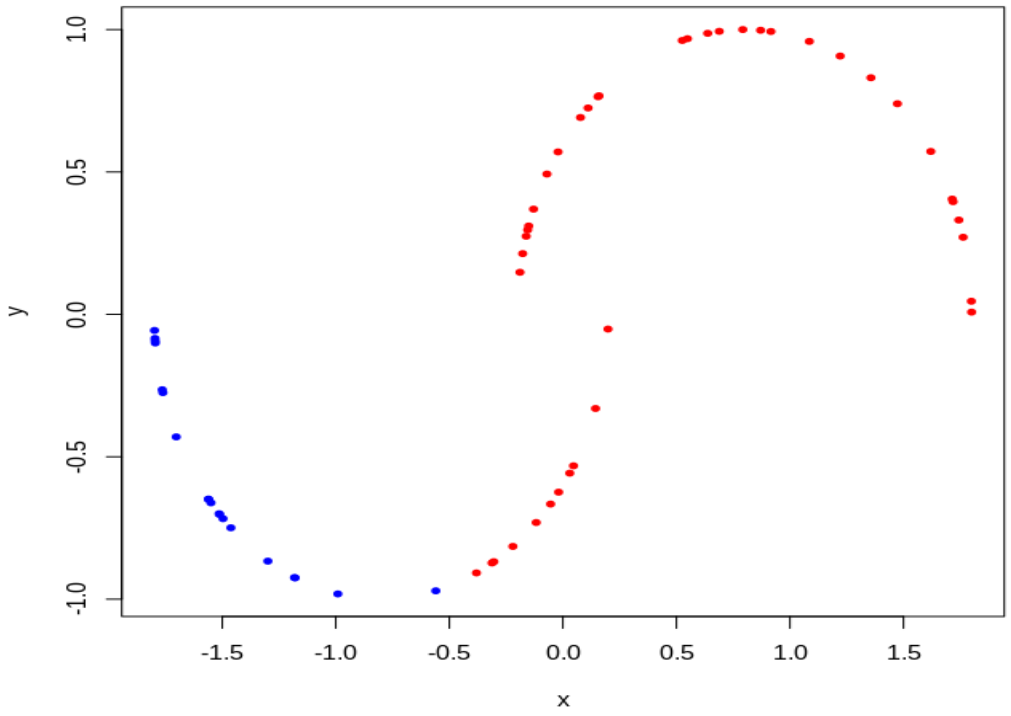


Figura 17: Visualización del agrupamiento en los primeros 2 eigenvectores menores en el conjunto de datos original. Para $\delta = 15$

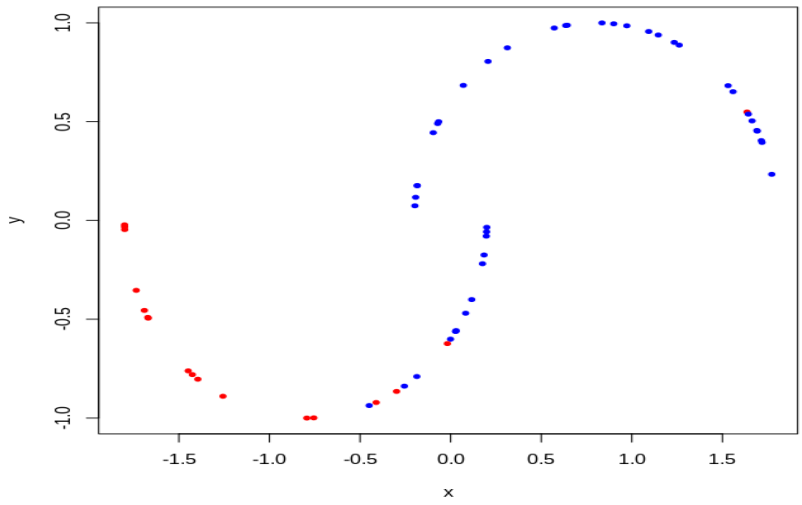
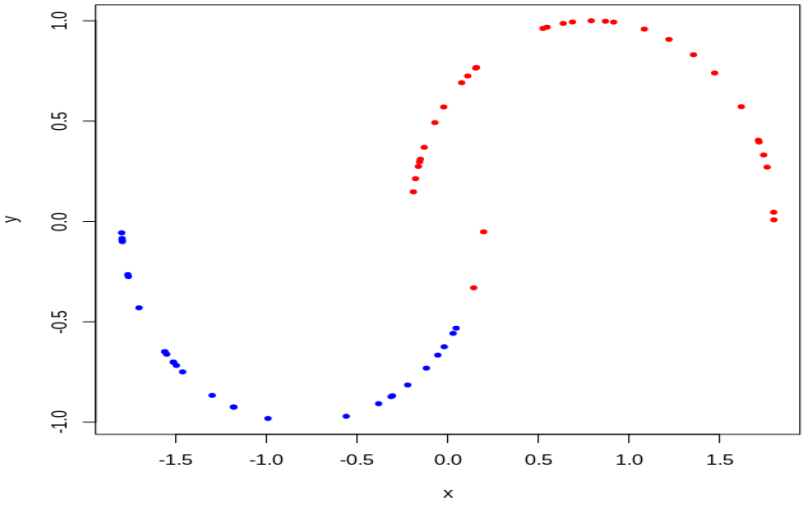


Figura 18: Visualización del agrupamiento en los primeros 2 eigenvectores menores en el conjunto de datos original. Para $\delta = 2$. (Mejores resultados)



Estas variaciones se pueden explicar si observamos los levelplot de las matrices de similitud correspondientes. Como se observa a continuación, aunque en todos los casos ilustrados, en dichas matrices aparecen dos grupos, el mayor contraste se observa exactamente para $\delta = 2$, mientras que en los casos extremales se difuminan.

Figura 19: Levelplot de matriz de similitud para $\delta = 2$

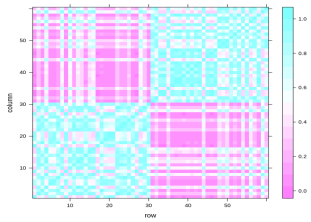


Figura 20: Levelplot de matriz de similitud para $\delta = 0.2$

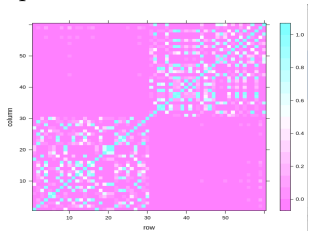
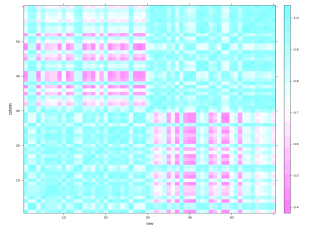


Figura 21: Levelplot de matriz de similitud para $\delta = 15$



Esta es, una vez más, una muestra de que para el reconocimiento de patrones la elección de las distancias es básica. En particular, en este ejemplo no se logró encontrar una distancia óptima (que coloreara a las lunas con colores distintos) debido a que si δ aumentaba, la región central tendía a colapsar, mientras que si se se disminuía, todos los puntos iban tendiendo a estar a la misma "distancia". Aún así, los resultados para ciertos valores de δ fueron buenos y coherentes, considerando que los conjuntos son convexos y cercanos.

Finalmente, cabe mencionar que el paso de la graficación del WGSS vs. K sólo se hizo de forma informativa pues para todos los casos, se eligió $K = 2$. Pero desde ese punto se sospechaba que algunos de los valores de δ iban a tener comportamientos erráticos puesto que para valores extremos la inclinación de la gráfica tendía a ser horizontal hasta valores altos (como 4 y 8). Mientras que en casos cercanos al 2, sí se advierten dos grupos.

Figura 22: WGSS vs. K, de K-medias. Para $\delta = 15$

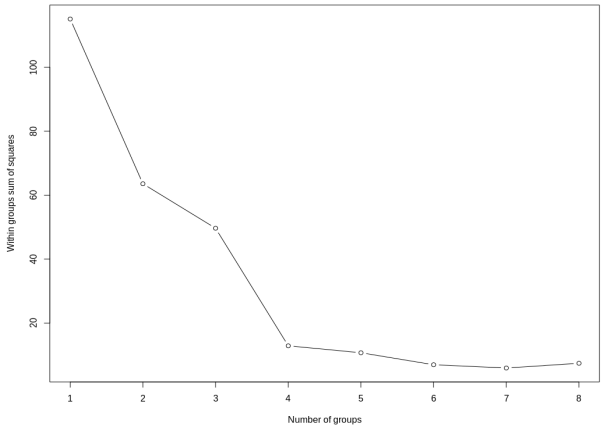


Figura 23: WGSS vs. K, de K-medias. Para $\delta = 2$