Tarea 4 Reconocimiento de Patrones

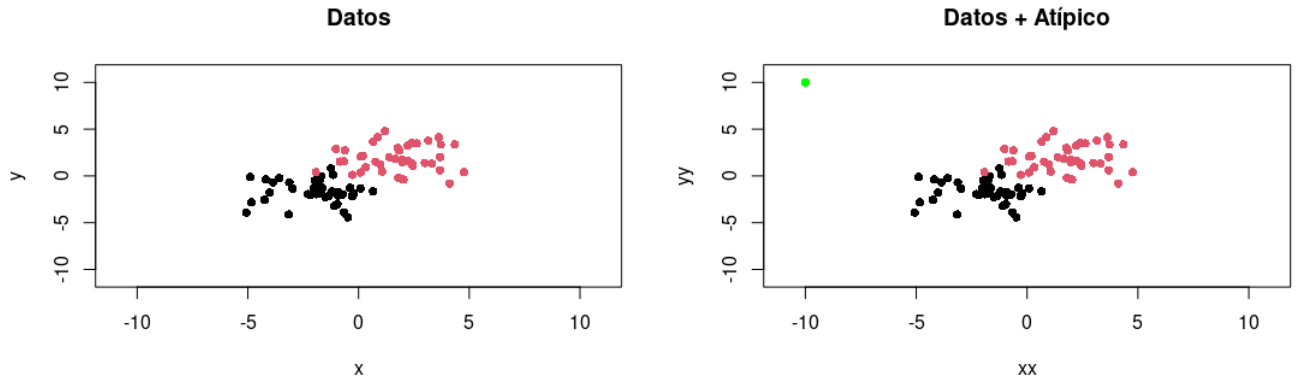
Yamil E. Morfa Avalos

7 de marzo de 2021

B. Preguntas cortas

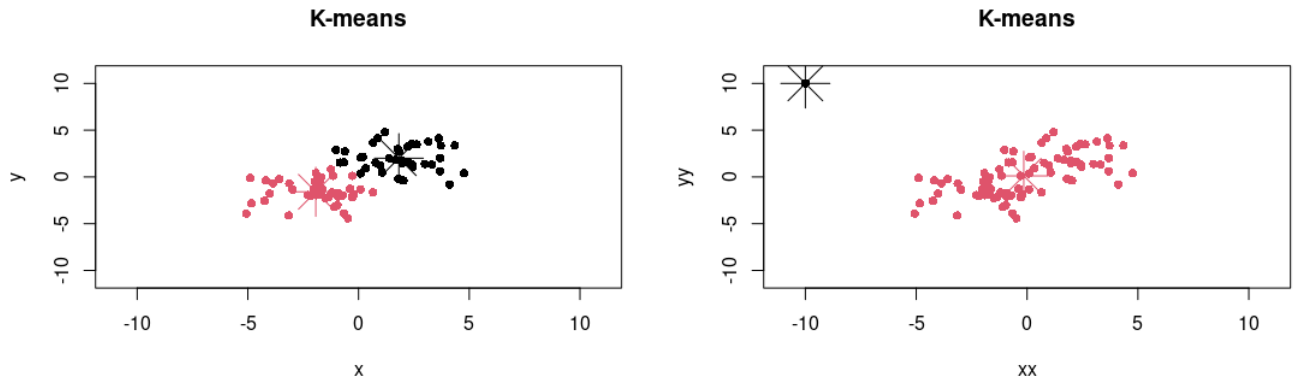
1. Entre los métodos de agrupamiento, k-medias o agrupamiento jerárquico, el más sensible a datos atípicos es k-medias.

Figura 1:



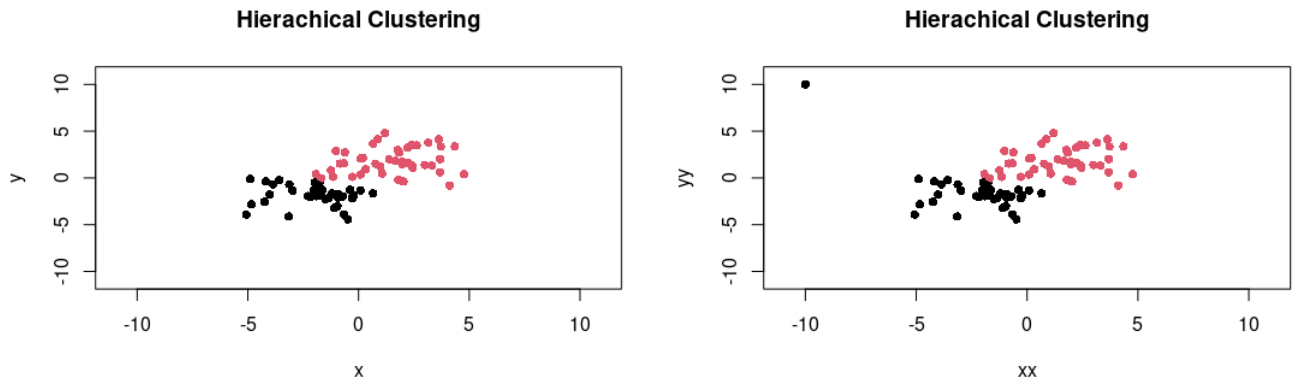
Supongamos los datos como aparecen en la figura 1. Dos normales bivariadas centradas en $\mu_1^T = [-2, -2]$ y $\mu_2^T = [2, 2]$ con $\Sigma_1 = \Sigma_2 = \sigma I$, para $\sigma = 1{,}5$. A estos se le han agregado un dato atípico $x^* = (-10, 10)$. Podemos justificar la hipótesis anterior sobre k-medias dado que al ser seleccionados los centros iniciales de manera aleatoria hay una posibilidad de que suceda lo que se muestra en la figura 2

Figura 2:



Mientras que con agrupamiento jerárquico siempre obtendremos el mismo resultado, como en la figura 3

Figura 3:



Veamos que con agrupamiento jerárquico no cambia la estructura de los grupos a pesar de la introducción del dato atípico.

2. Supongamos que $X_i,\, Y_i \sim Z$ i.i.d con $E(Z) = 0$ y $Var(Z) < +\infty$. Supongamos qué $X_i Y_i \sim \hat{Z}$, luego $E\left(\hat{Z}\right) = E^2(Z) = 0$ y $Var\left(\hat{Z}\right) < +\infty$. Tenemos entonces que:

$$\langle X, Y\rangle = \frac{1}{d}\sum_{i=1}^{d} X_i Y_i = \overline{\hat{Z}}_d$$

Luego por la ley débil de los grandes números, para cualquier $\varepsilon > 0$ se tiene que:

$$\lim_{d\to+\infty} P\left(\left|\overline{\hat{Z}}_d - E\left(\hat{Z}\right)\right| \leq \varepsilon\right) = 1 \Longrightarrow \lim_{d\to+\infty} P\left(|\langle X, Y\rangle| \leq \varepsilon\right) = 1$$

luego:

$$\lim_{d\to+\infty} P\left(|\langle X, Y\rangle| > \varepsilon\right) = \lim_{d\to+\infty} \left(1 - P\left(|\langle X, Y\rangle| \leq \varepsilon\right)\right) = 1 - \lim_{d\to+\infty} P\left(|\langle X, Y\rangle| \leq \varepsilon\right) = 0$$

C. Análisis de datos

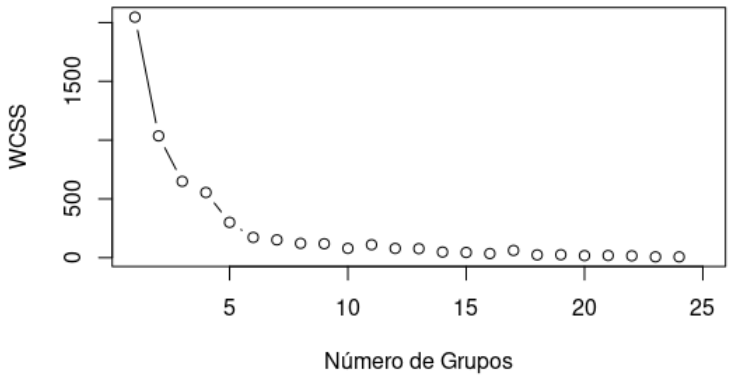
1. Para los datos de heptatlon que están en *library* $('MVA')$ donde se muestran los resultados de las 7 pruebas de heptatlon para 25 atletas. La ultima columna que representa las marcas según el sistema de puntuación del heptatlon no serán consideradas para el agrupamiento.

En la figura 4 se muestra para cada valor de $k \in [1, 2, ..., 25]$ los valores de $WCSS$

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

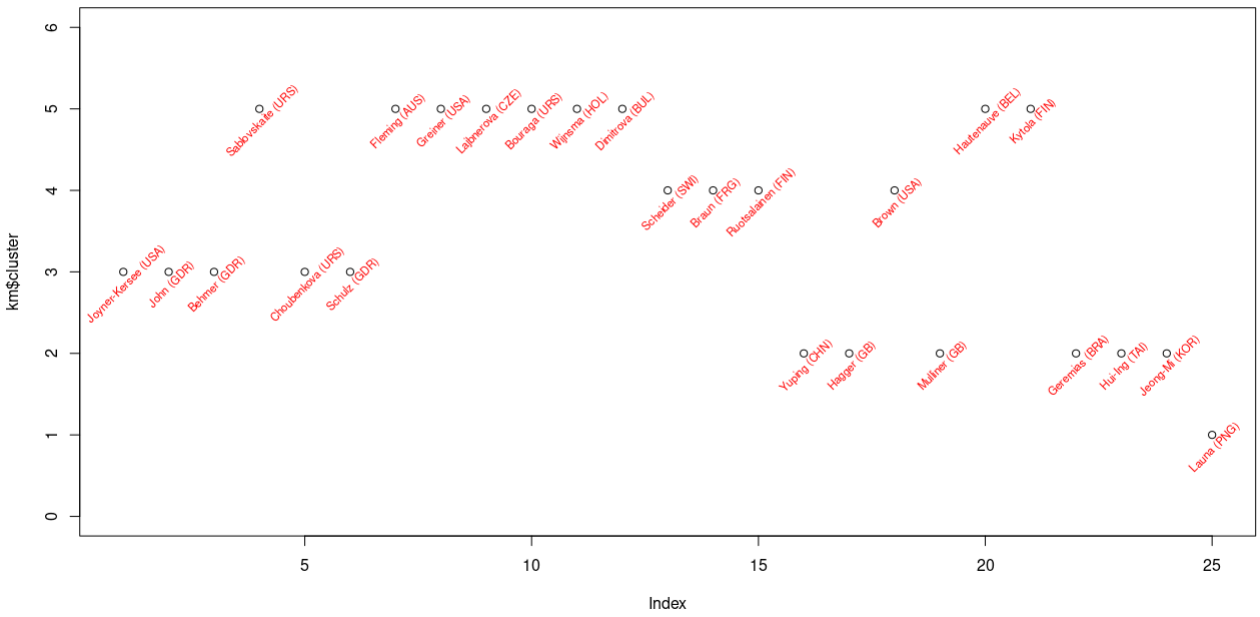
con S_i $i = 1, 2, .., k$ los k grupos obtenidos con centros μ_i $i = 1, 2, .., k$ y $x_j \in \mathbb{R}^7$ las observaciones.

Figura 4:



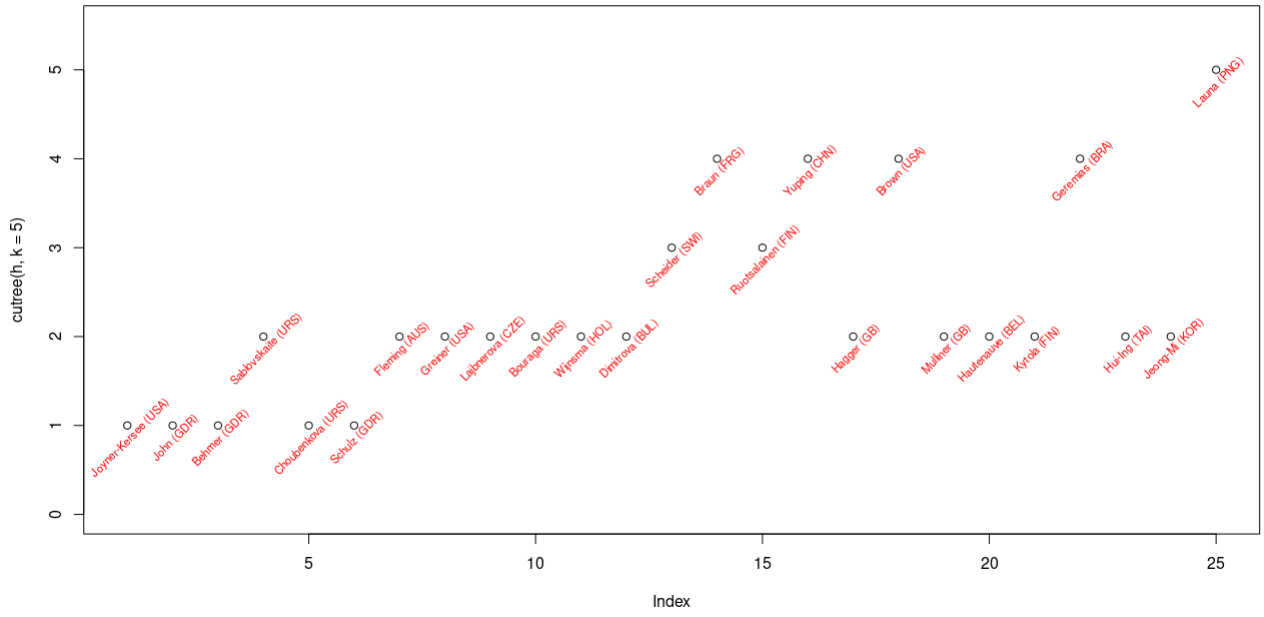
Podemos notar que aunque el mínimo valor de estos se obtiene para $k = 25$ no tiene sentido hacer este agrupamiento puesto que tenemos 25 datos. Sin embargo podemos ver que el ultimo cambio en los valores de $WCSS$ que podría resultar relevante es el cambio a 5 . Así que para este ejemplo se ha seleccionado $k = 5$.

Figura 5:



En la figura 5 se muestran como quedan agrupado los aletas en los 5 grupos. Veamos que para *Launa* (PNG) la marca obtenida es 4566, la cual es considerablemente menor que las marcas obtenidas por los demás atletas y en consecuencia el primer grupo solo consiste de este elemento. También podemos observar que los 6 primeros lugares según la puntuación, 5 están en el mismo grupo. Solo Sablovskaite (URS) quien obtuvo el puesto 4 aparece agrupado junto a los atletas que obtuvieron del $7mo$ al $12mo$ lugar (¿podría este atleta haber sido mal puntuado?).

Figura 6:



En la figura 6 se muestra sucede similar pues en un grupo se encuentran 5 de los primeros 6 lugares y la excepción también en Sablovskaite (URS), además también se le asigna un grupo separado a *Launa* (PNG). Parece ser que la manera de agrupar estos datos tiene cierta relación con la manera en la que se asignan las puntuaciones.

2. (Compañera de equipo Lilian Bárbara Pérez Sosa)

Considerando los datos del proyecto de la Universidad de Oxford sobre las diferentes medidas que los gobiernos tomaron para enfrentar COVID-19: https://covidtracker.bsg.ox.ac.uk/ , se han tomado los datos correspondientes al 01/01/2021 para todos los países solo guardando las variables referentes a C1_School closing, C2_Workplace closing, C3_Cancel public events, C4_Restrictions on gatherings, C5_Close public transport, C6_Stay at home requirements, C7_Restrictions on internal movement, C8_International travel controls. Además se han eliminado las observaciones donde en alguna de estas variables no se tenía datos.

A modo de hipótesis podemos suponer que las medidas para ciertos países de América del Norte, Sur América, Europa del Este y Oeste, África y Asia dado su cercanía y semejanza en su modo de vida, fuero de alguna forma similares. Teniendo esto en cuanta propondremos $k = 6$ grupos distintos.

Figura 7:

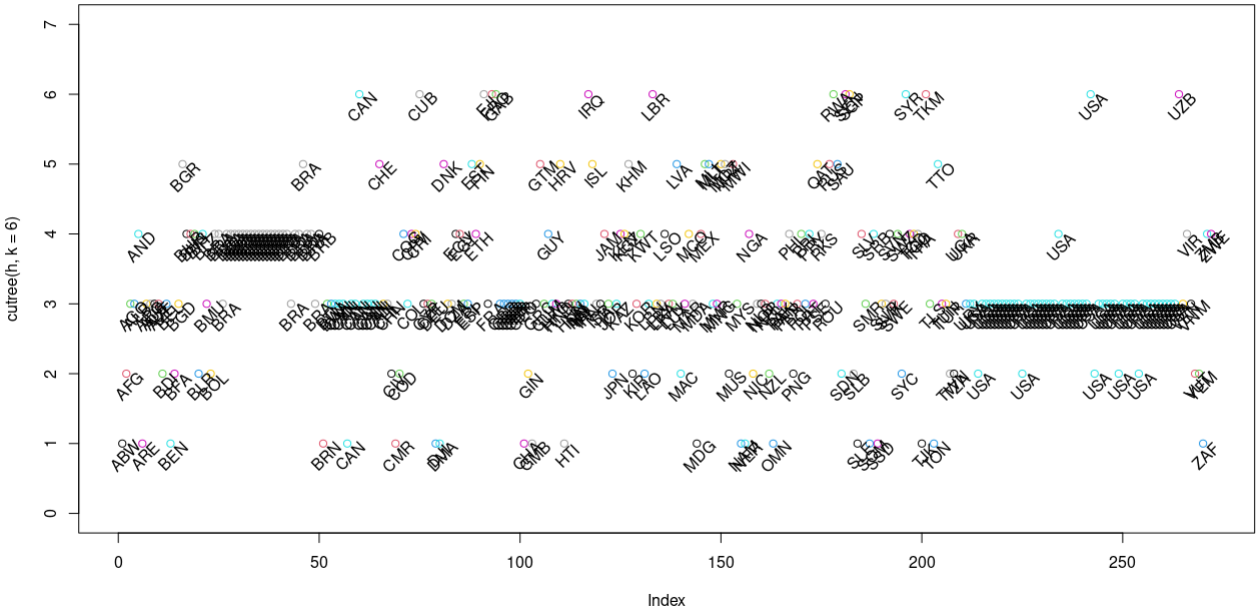
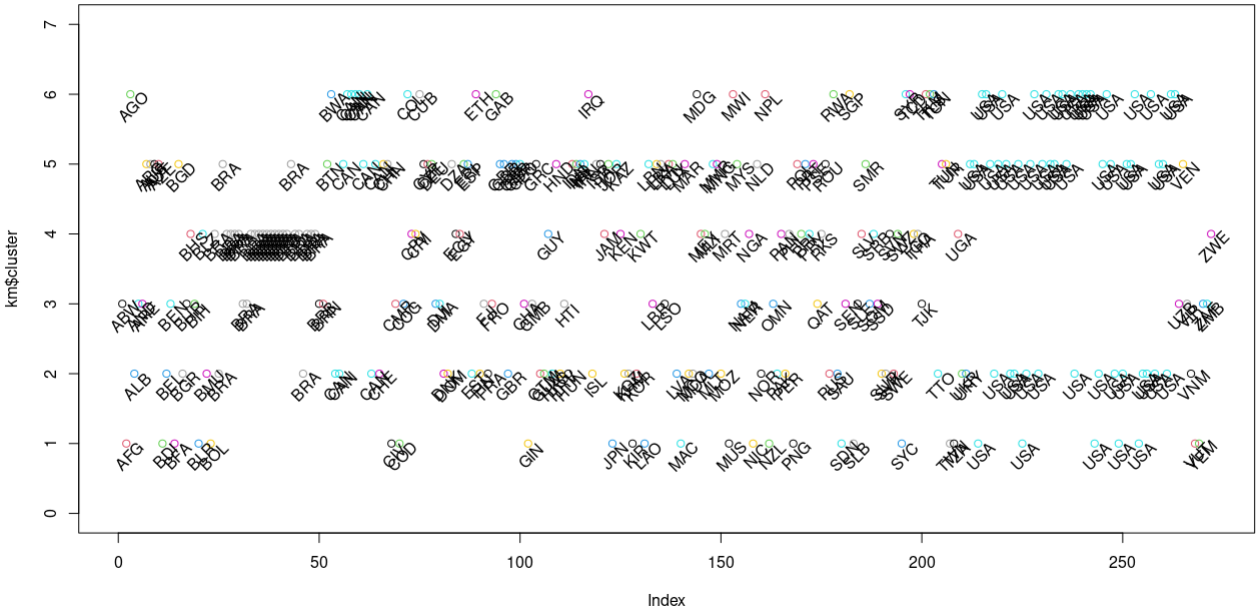


Figura 8:



En la figura 7 y 8 se muestran los grupos obtenidos para agrupamiento jerárquico y k-medias respectivamente. Como podemos ver, estos no siguen una estructura relacionada con la hipótesis supuesta pues en un mismo grupo podemos encontrar varios países de regiones diferentes. Podemos pensar que se deba a que el estudio se realizó solamente para un único día y para un pequeño conjunto de variables. Consideramos que se realizaría un mejor análisis si se tomara en cuenta el cambio de estas medidas en varios días y se relacionara esto de alguna forma con la situación de los países frente a la pandemia.