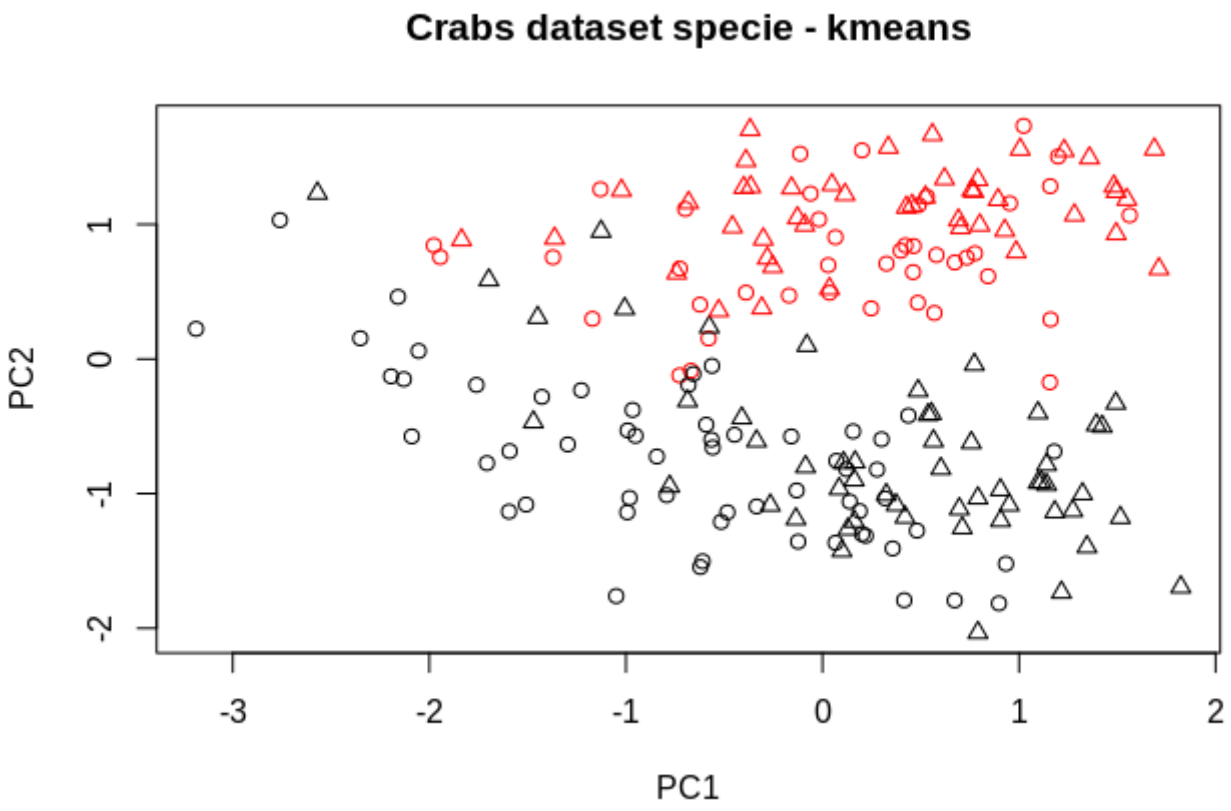
Trabajo Práctico 3: Clustering

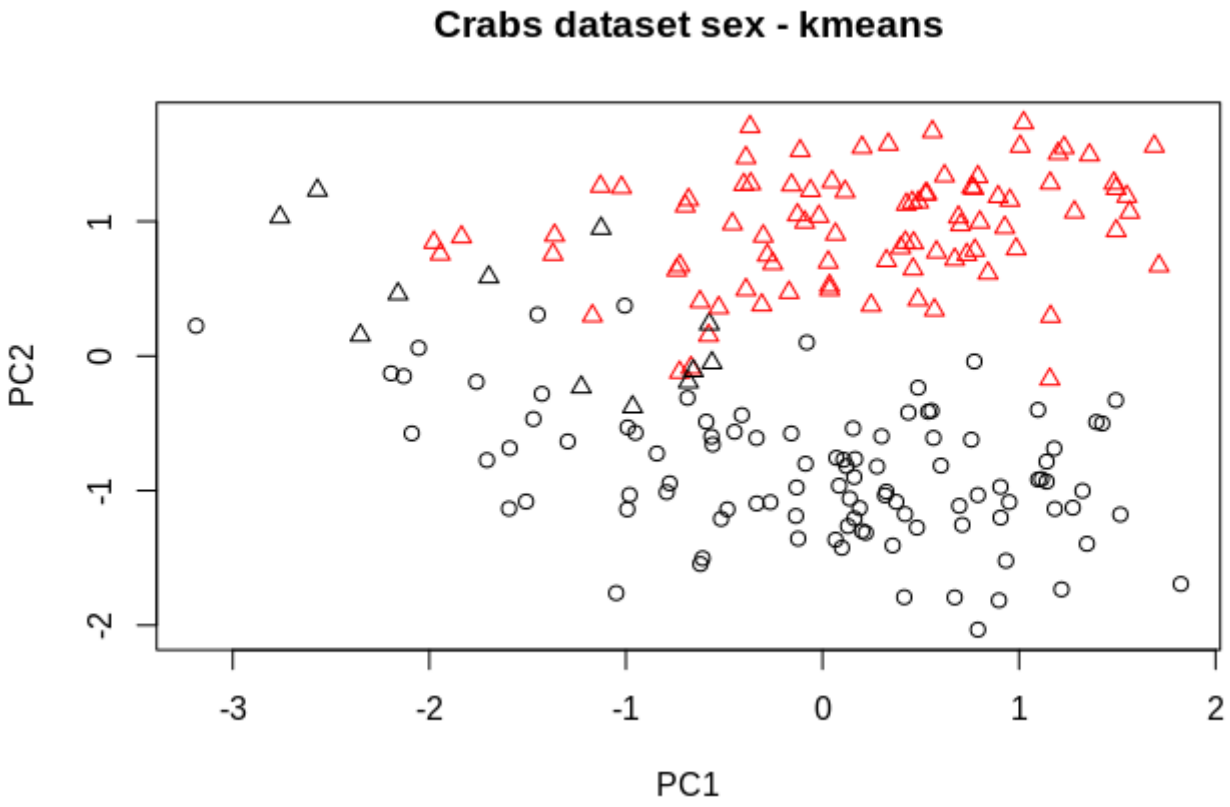
Minería de Datos

Alumno: Navall, Nicolás Uriel. N-1159/2.

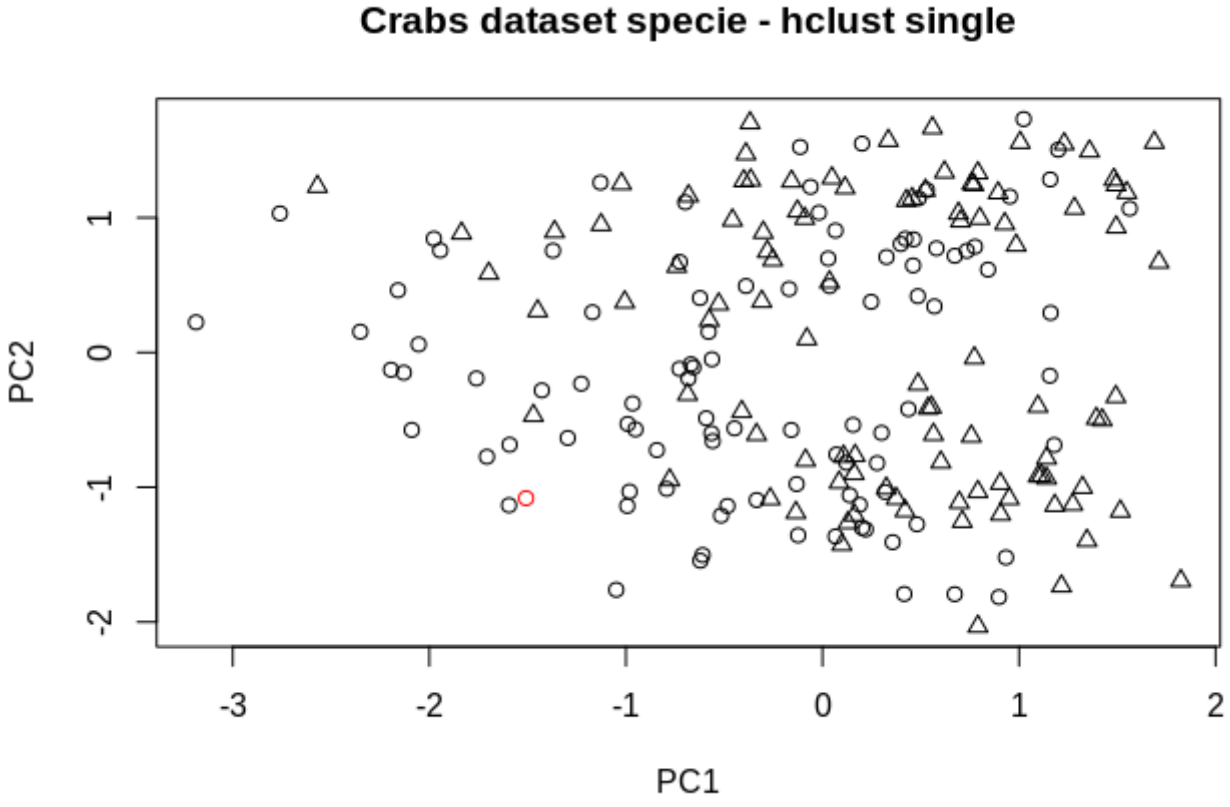
1)a) En las gráficas siguientes las formas de los puntos representan una variable del dataset (especificada en el título de cada gráfica) y los colores representan los clusters encontrados por el método utilizado (también especificado en el título de cada gráfica). Debajo de cada gráfica se encontrarán los resultados porcentuales de que tan bien supo encontrar la clasificación buscada el método utilizado.



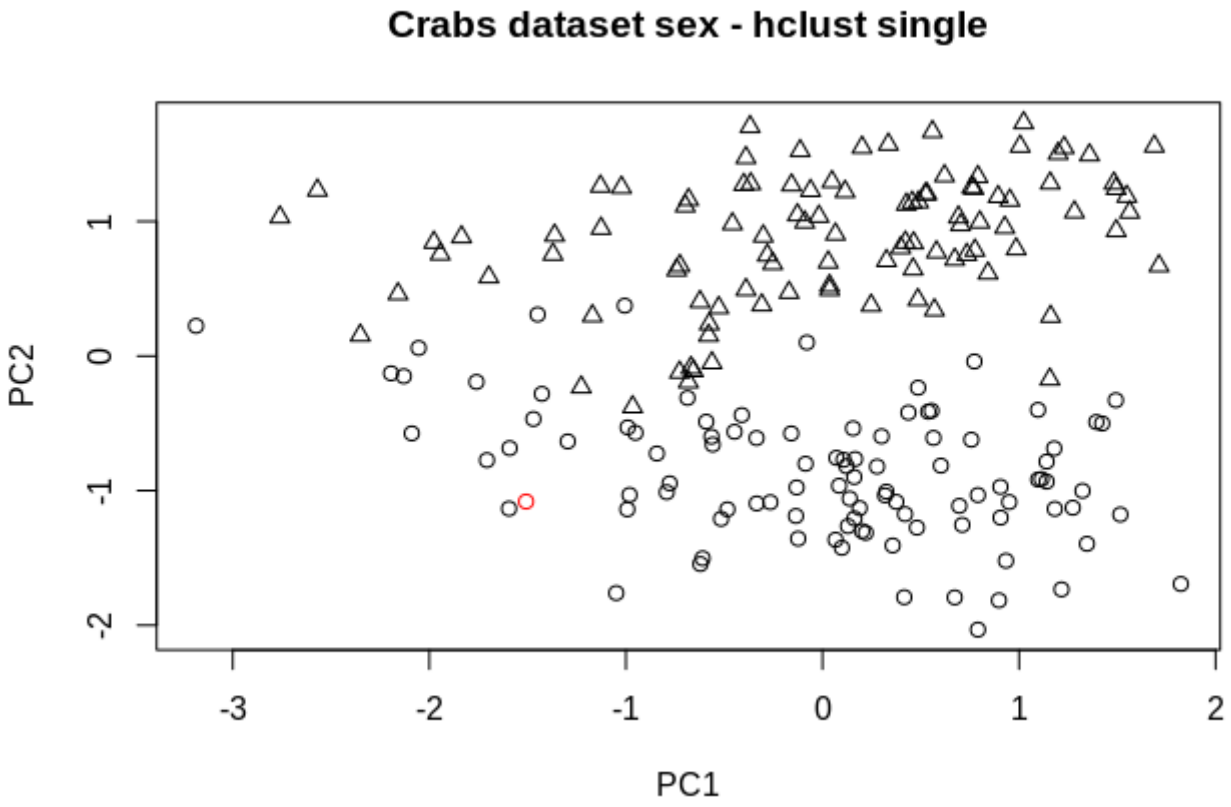
Direct agreement: 0 of 2 pairs
Iterations for permutation matching: 2
Cases in matched pairs: 53 %



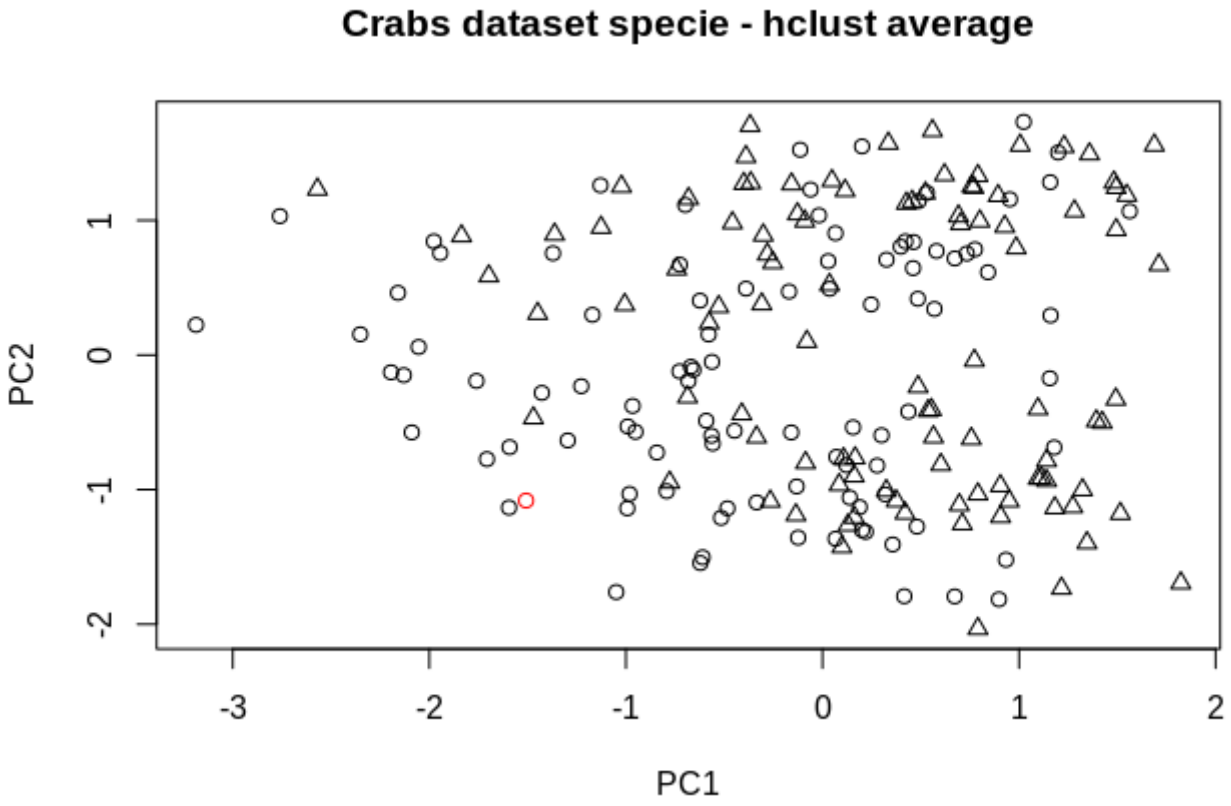
Direct agreement: 2 of 2 pairs
Cases in matched pairs: 93 %



Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 50.5 %

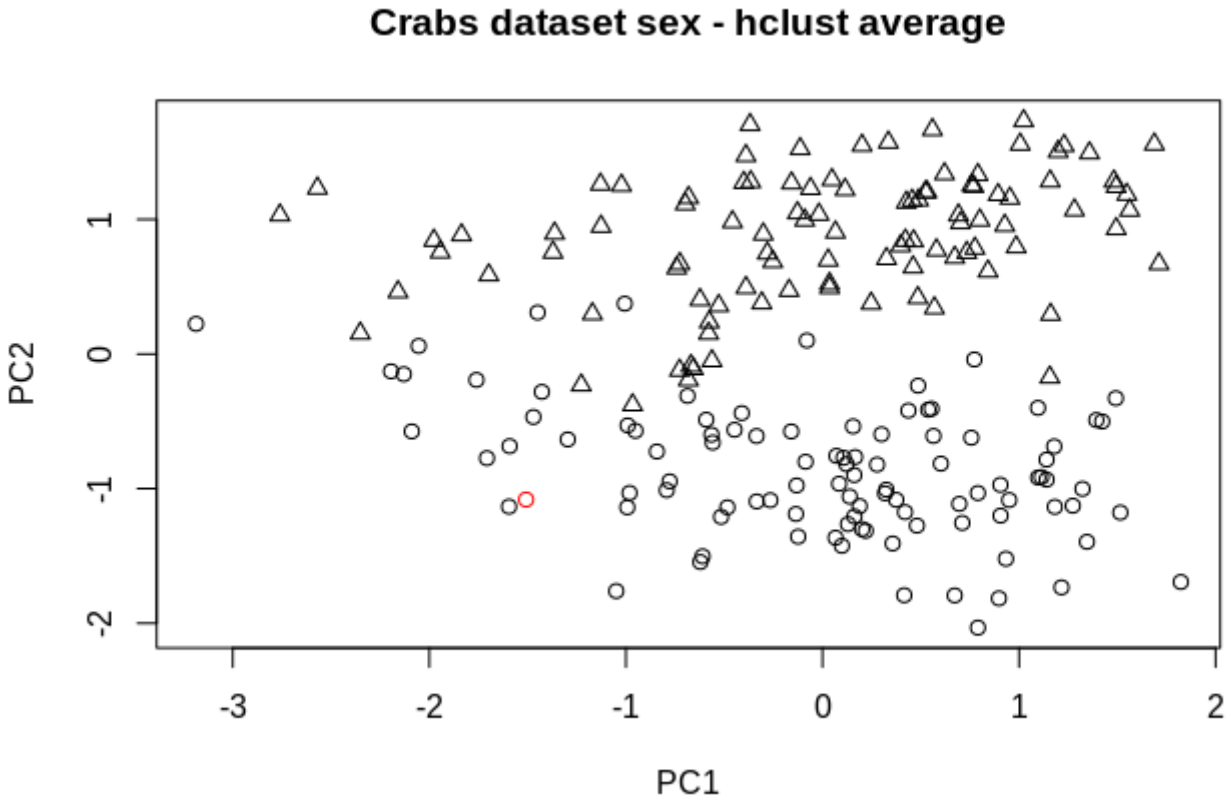


Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 50.5 %

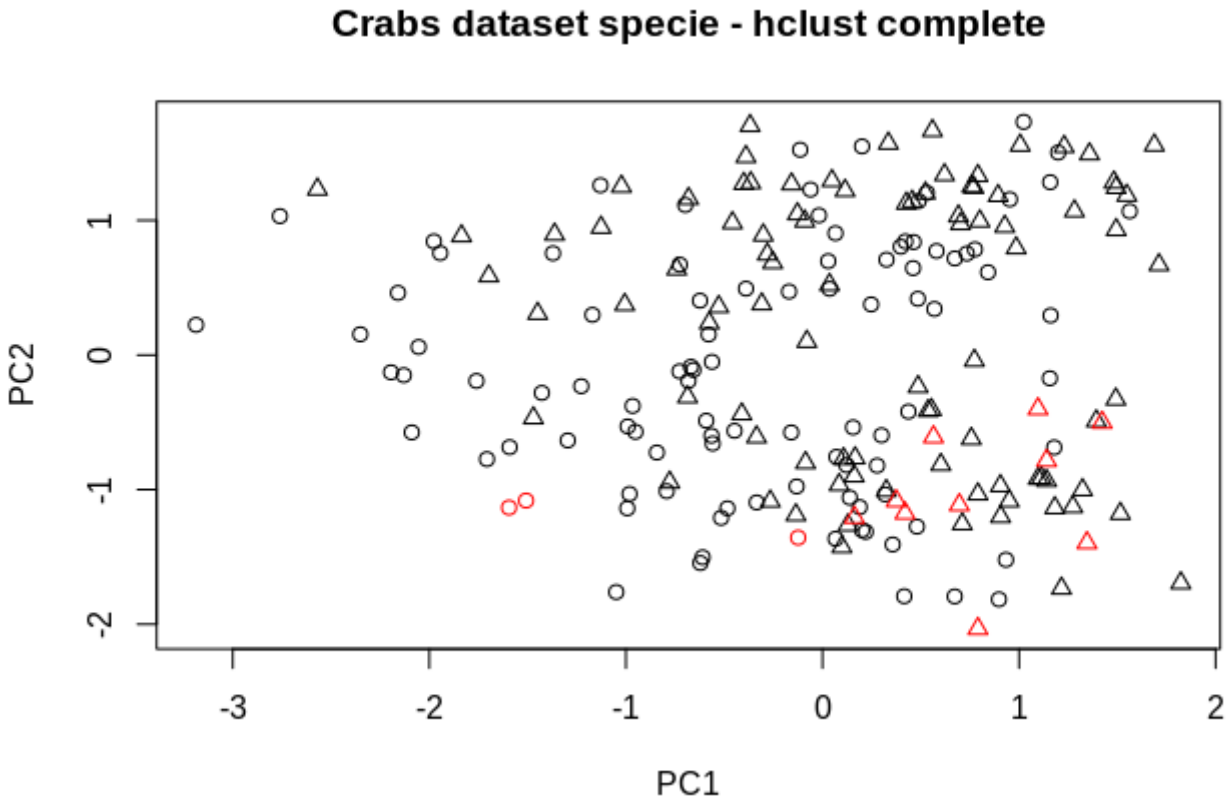


Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1

Cases in matched pairs: 50.5 %

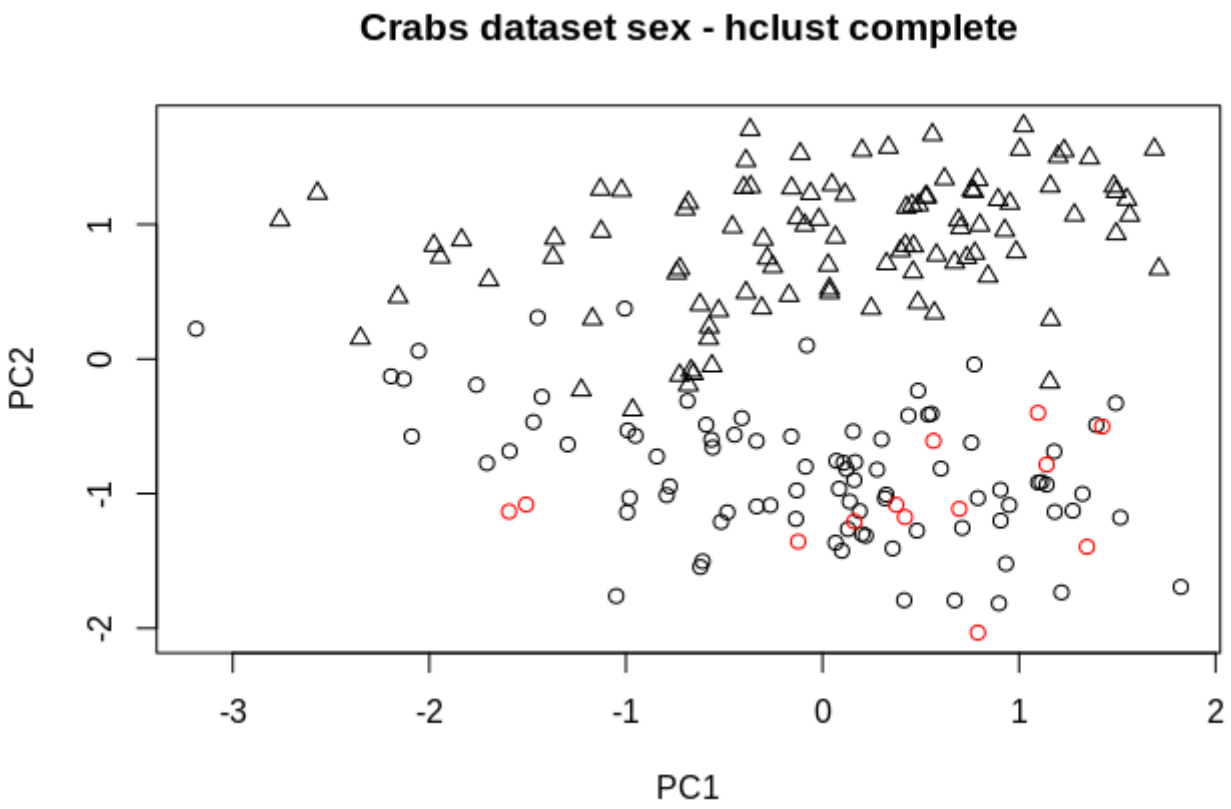


Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 50.5 %



Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1

Cases in matched pairs: 53.5 %



Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 56.5 %

Con los resultados obtenidos en kmeans podemos observar que al buscar dos clusters en el dataset estos reflejan la clasificacion por sexo del dataset, lo que nos da a entender que los datos medidos de los cangrejos son afectados en mayor medida por su sexo.

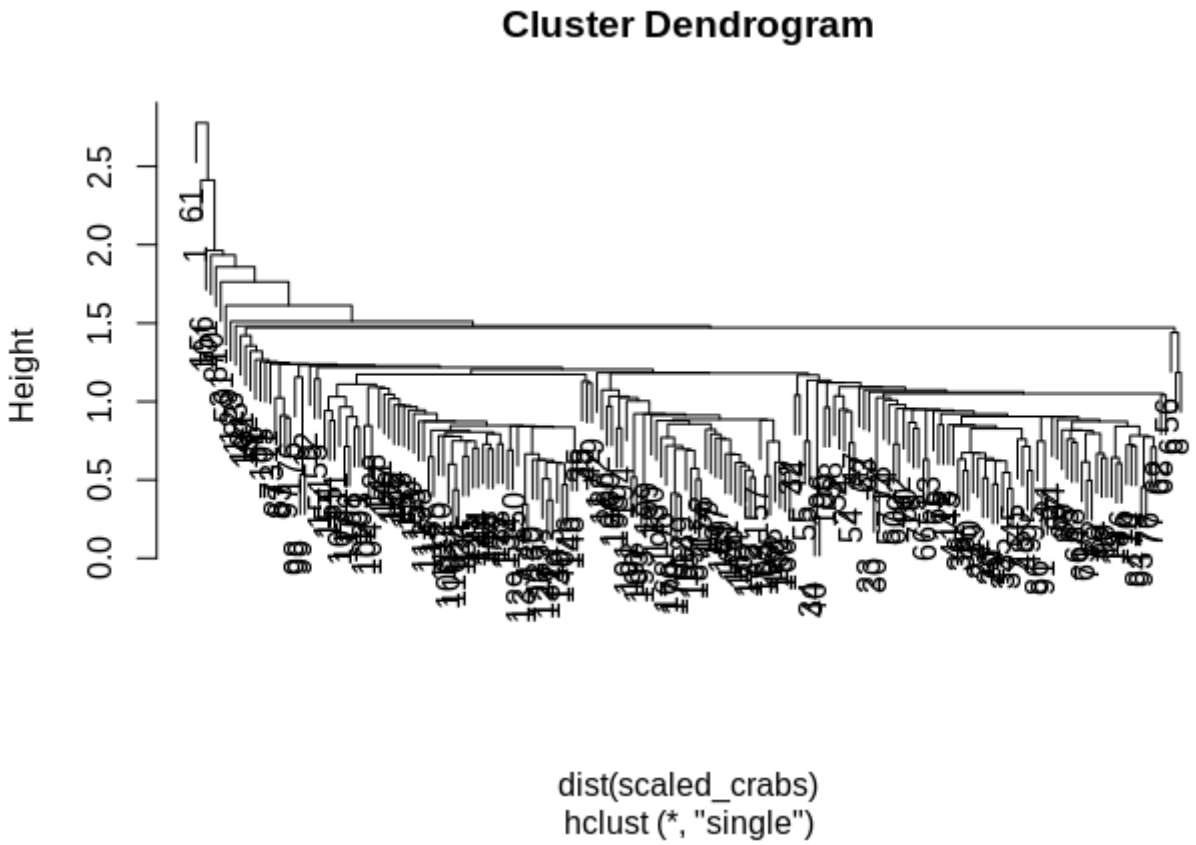


Figura 1.1: Dendrograma de hclust single sobre el dataset crabs transformado

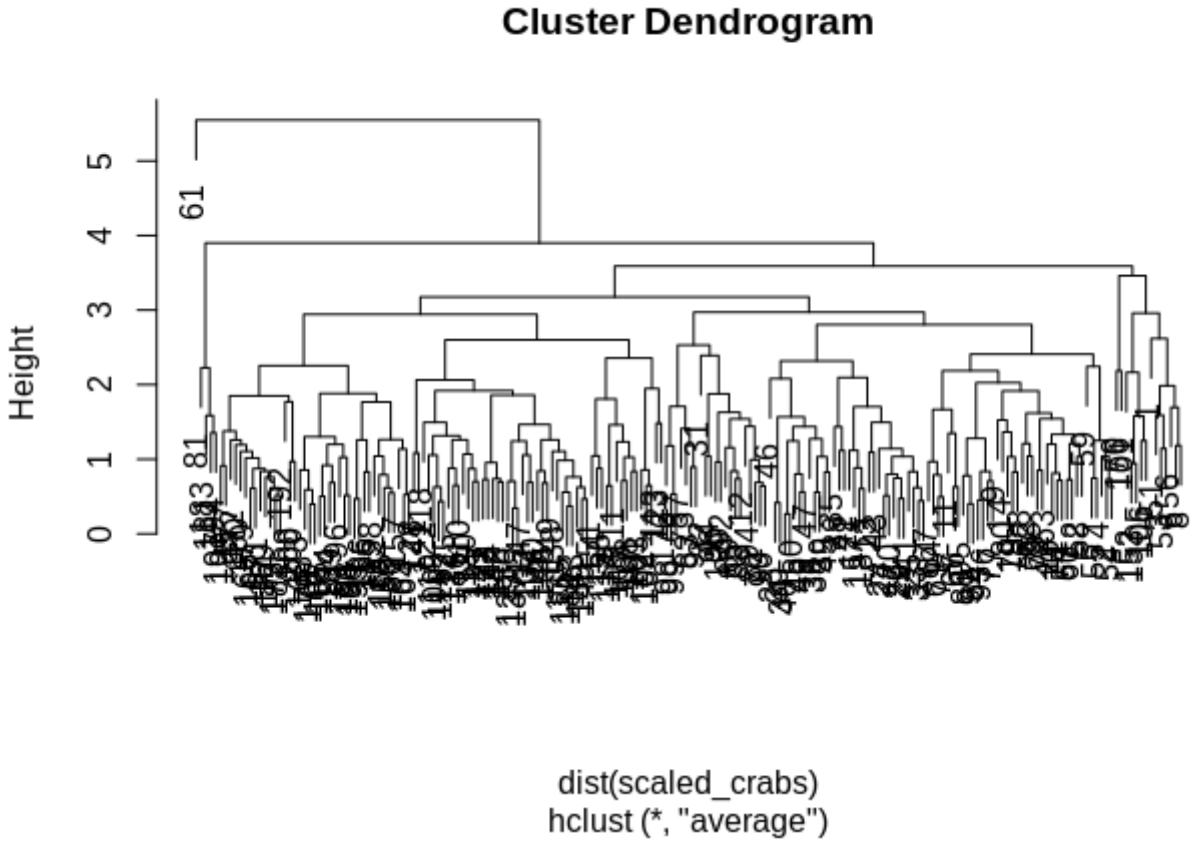


Figura 1.2: Dendrograma de hclust average sobre el dataset crabs transformado

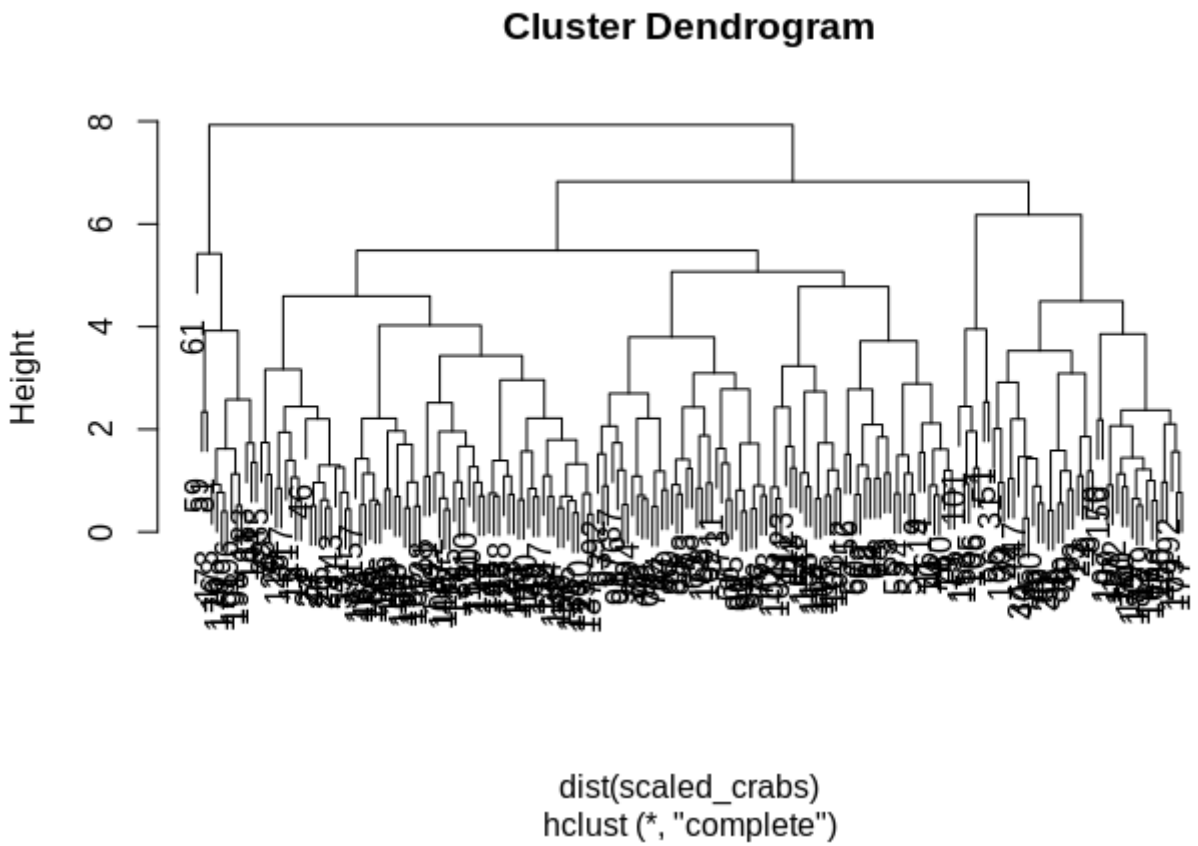
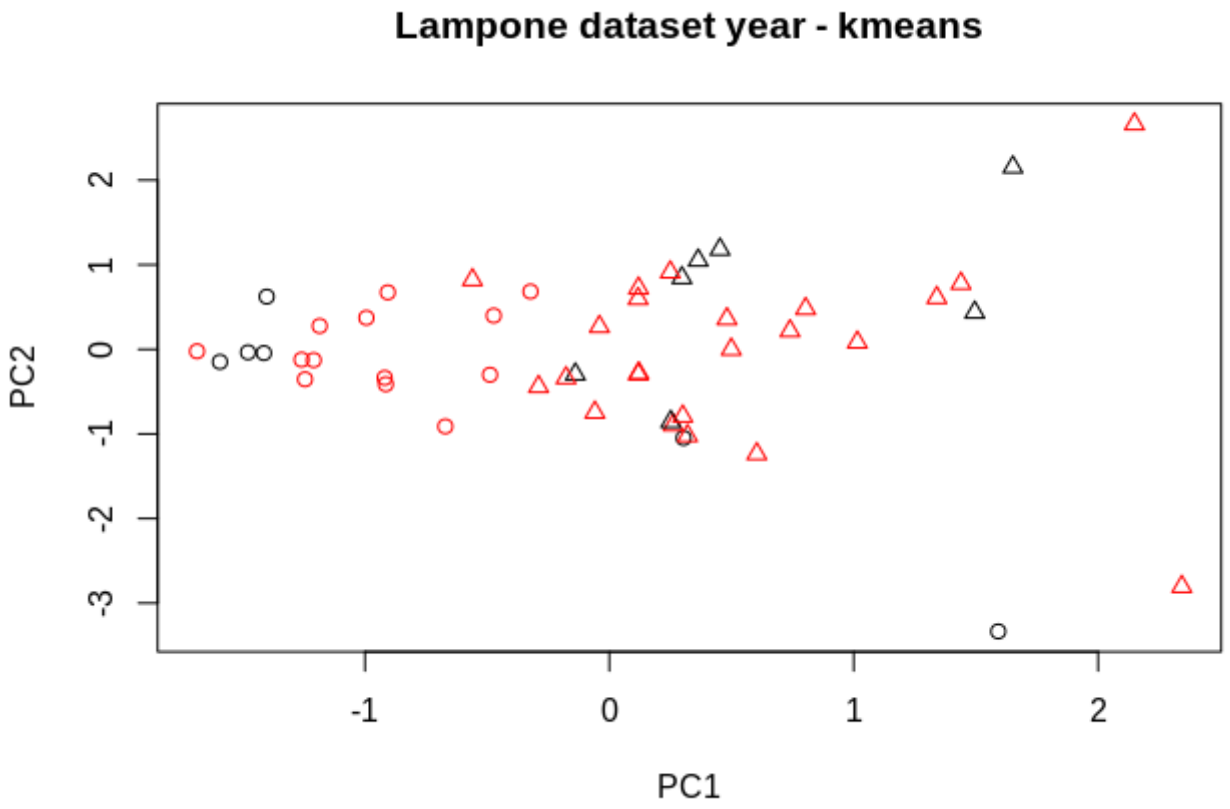


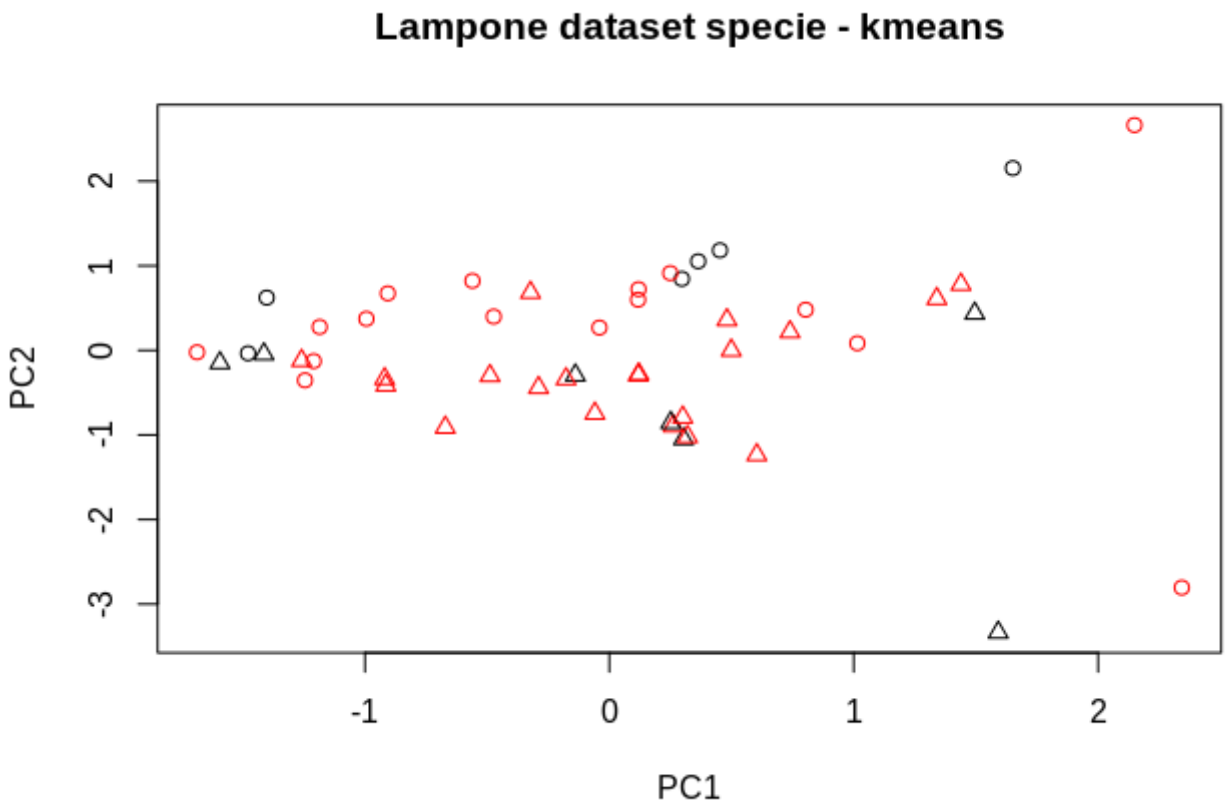
Figura 1.3: Dendrograma de hclust complete sobre el dataset crabs transformado

Como podemos observar en los dendrogramas, los métodos jerárquicos sobre este dataset no dan buenos resultados, ya que genera clusters a partir de outliers en las primeras ramas. Esto también lo podemos ver reflejado en las gráficas anteriores en donde uno de los clusters solo tenía un punto.

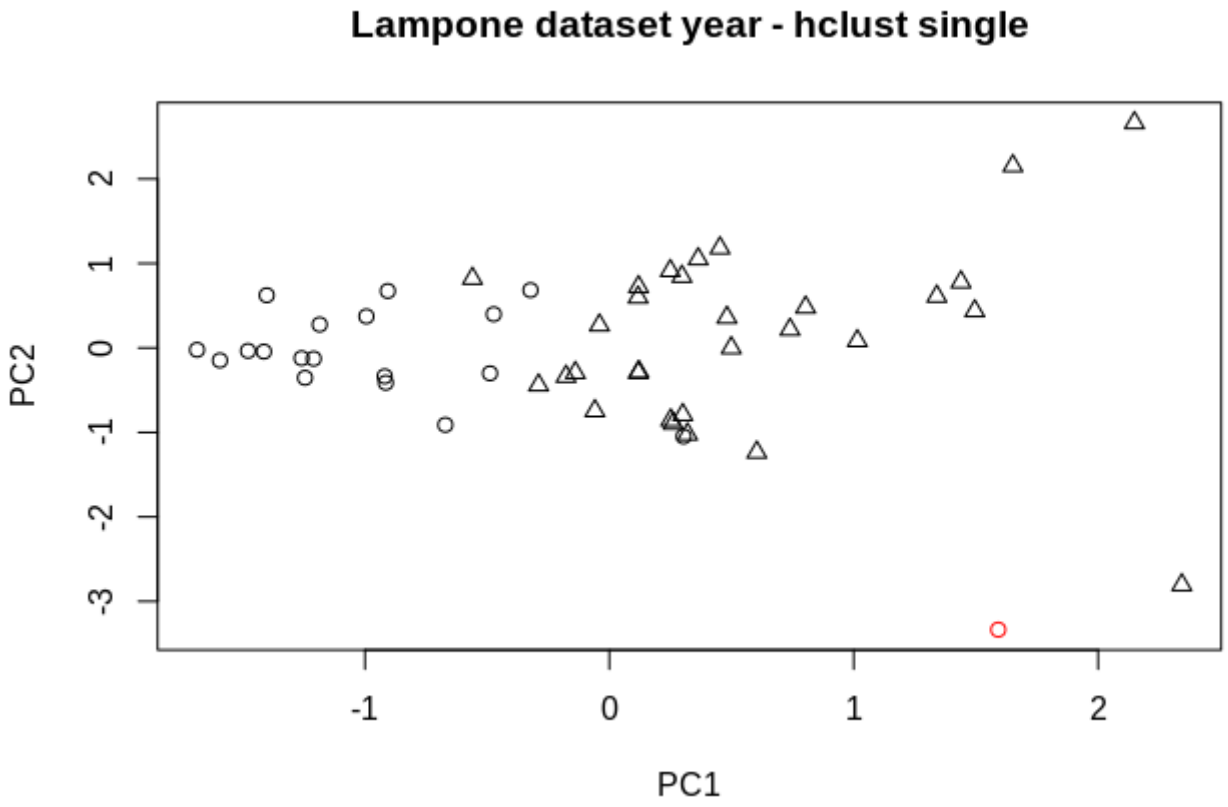
b)



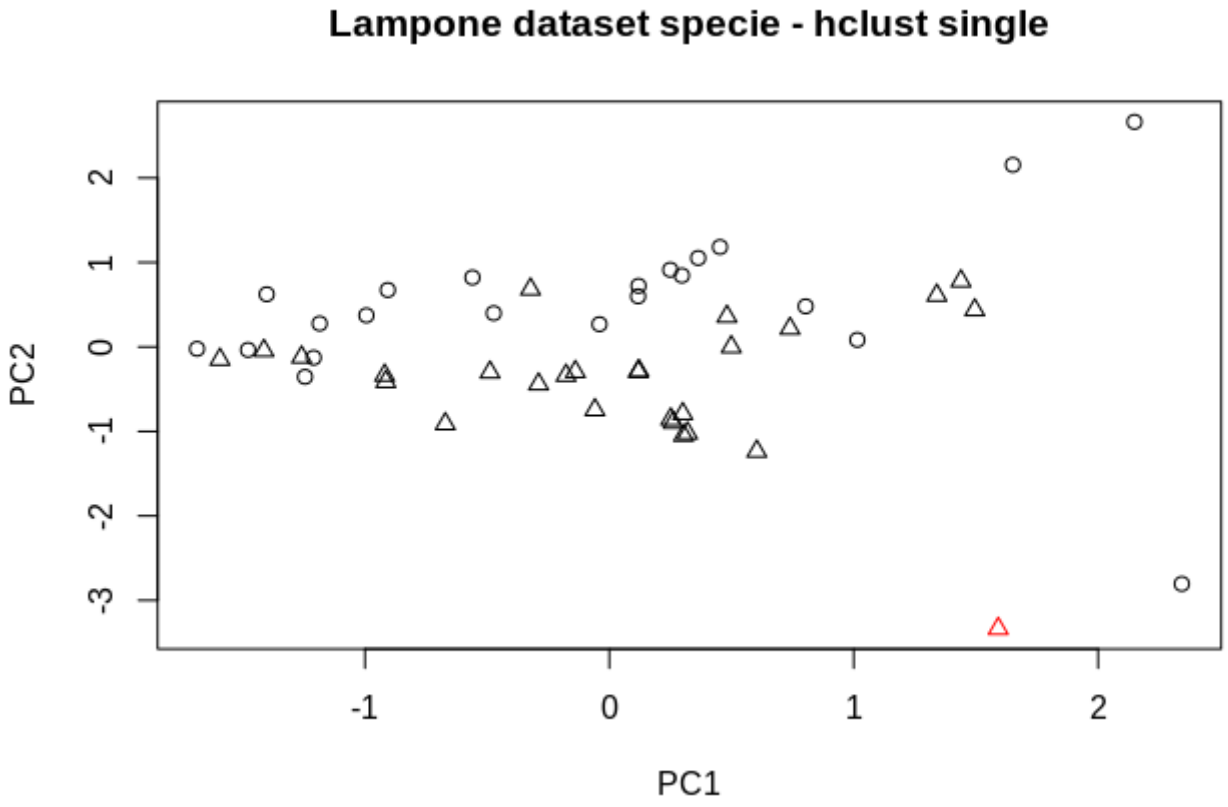
Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 59.18 %



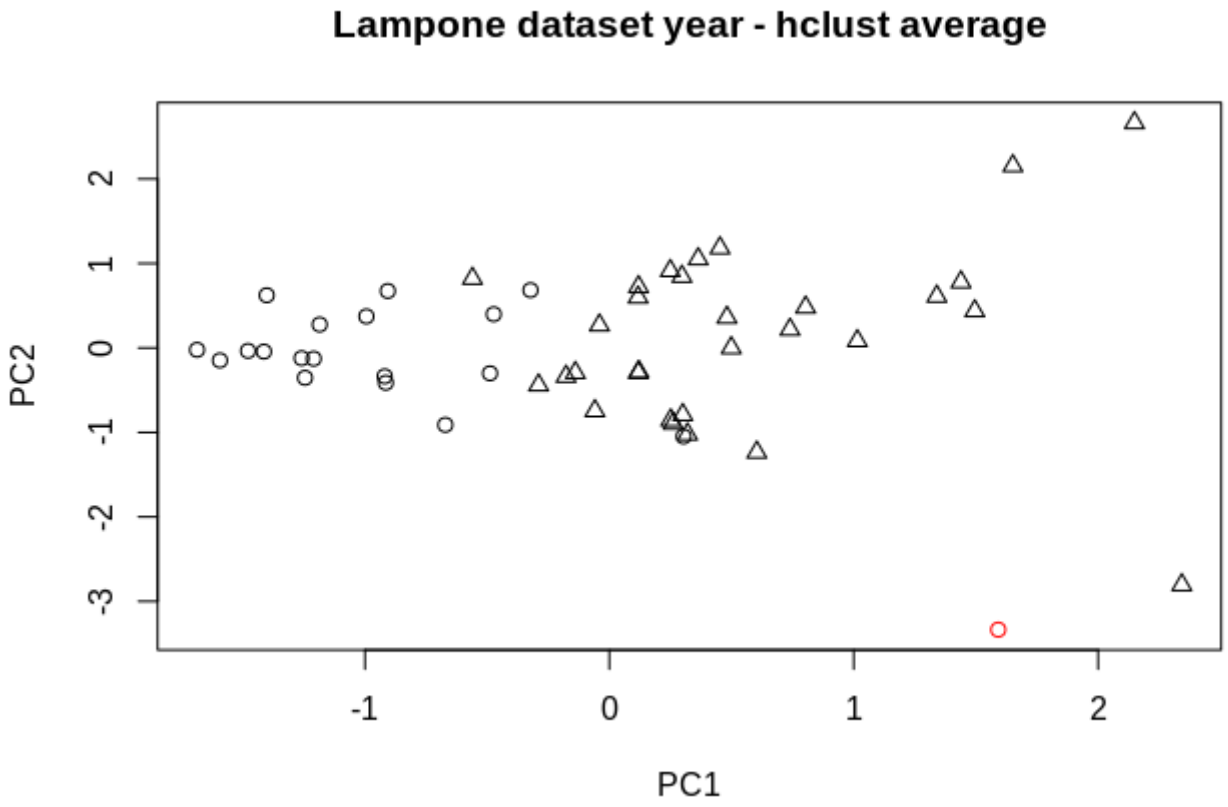
Direct agreement: 0 of 2 pairs
Iterations for permutation matching: 2
Cases in matched pairs: 53.06 %



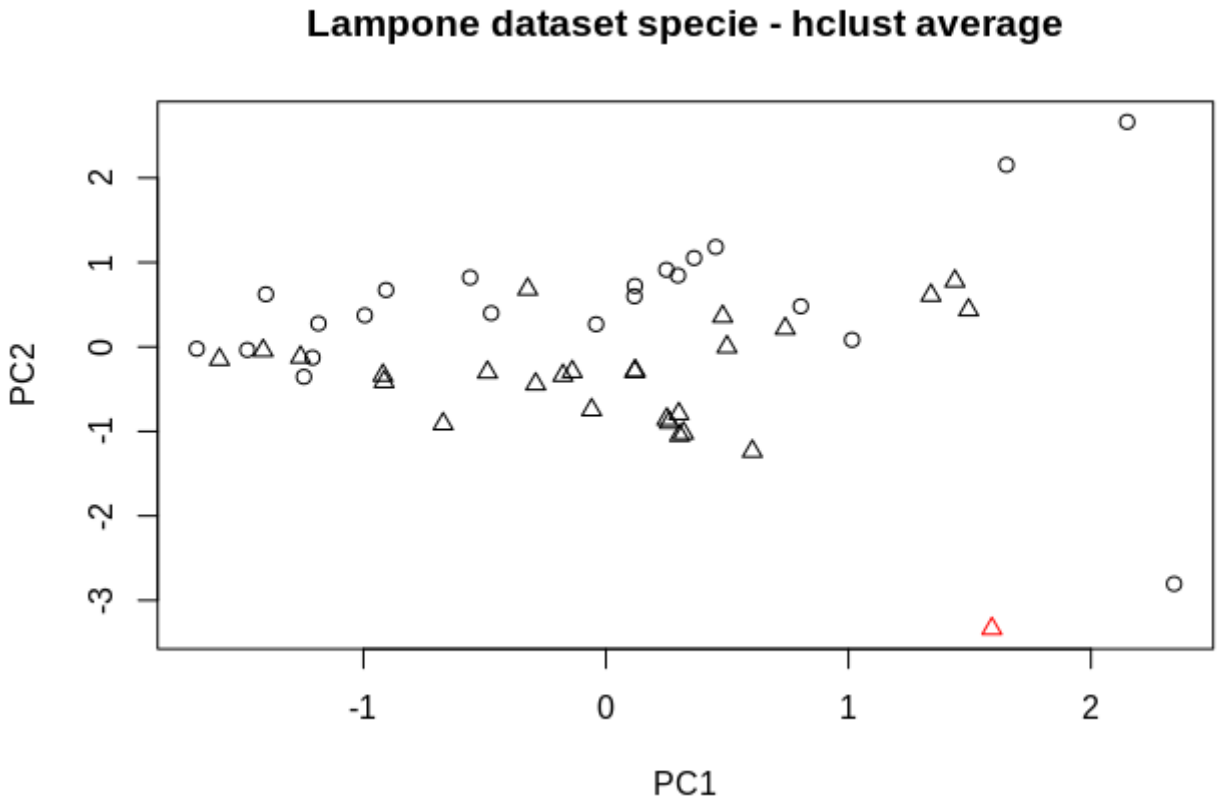
Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 63.27 %



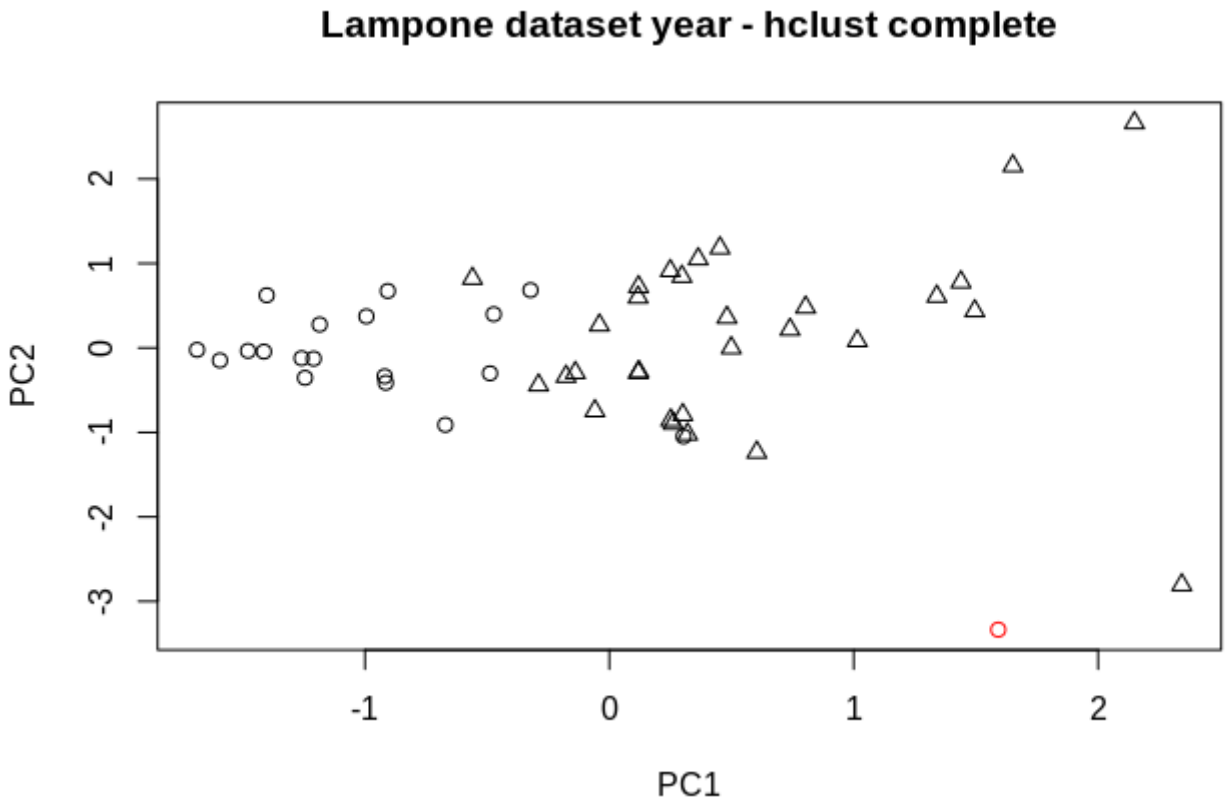
Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 53.06 %



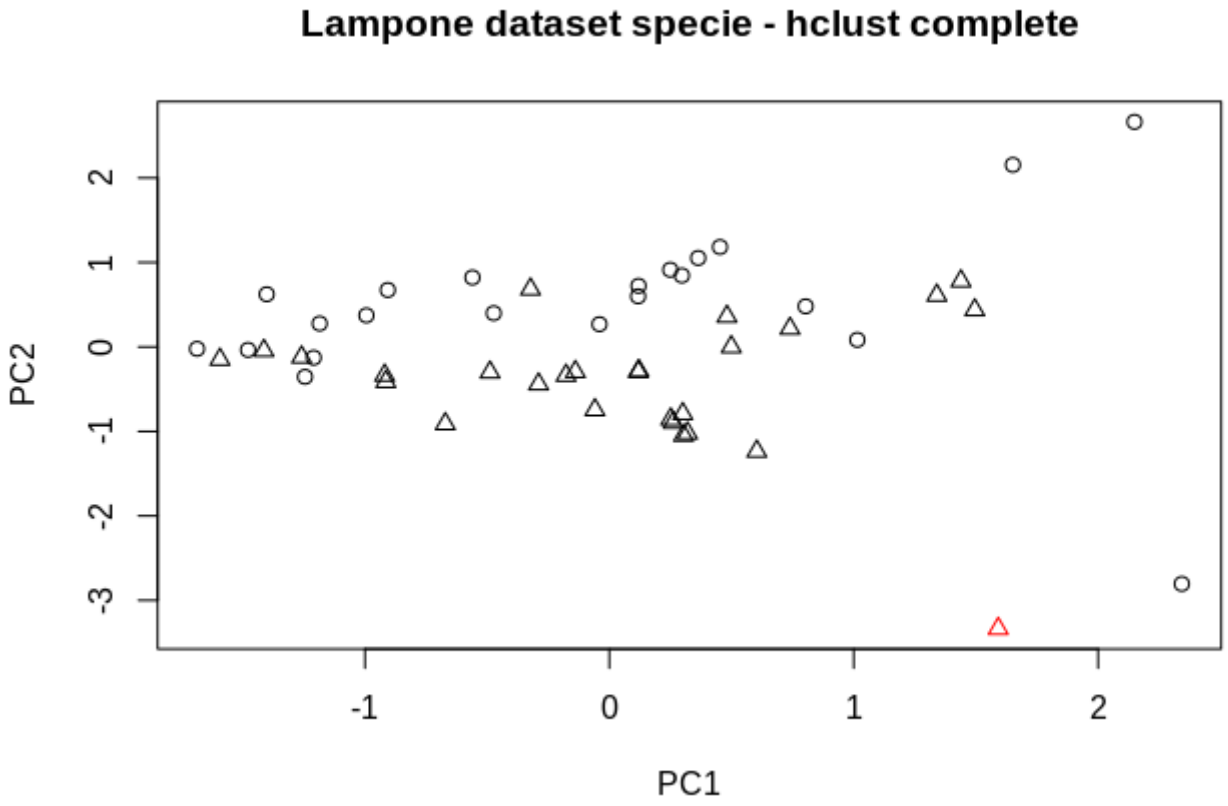
Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 63.27 %



Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 53.06 %



Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 63.27 %



Direct agreement: 1 of 2 pairs
Iterations for permutation matching: 1
Cases in matched pairs: 53.06 %

Resulta muy útil aplicar PCA a estos datos por la gran cantidad de variables que tiene, ya que será muy probable que no se usen todos estos para clasificarlos.

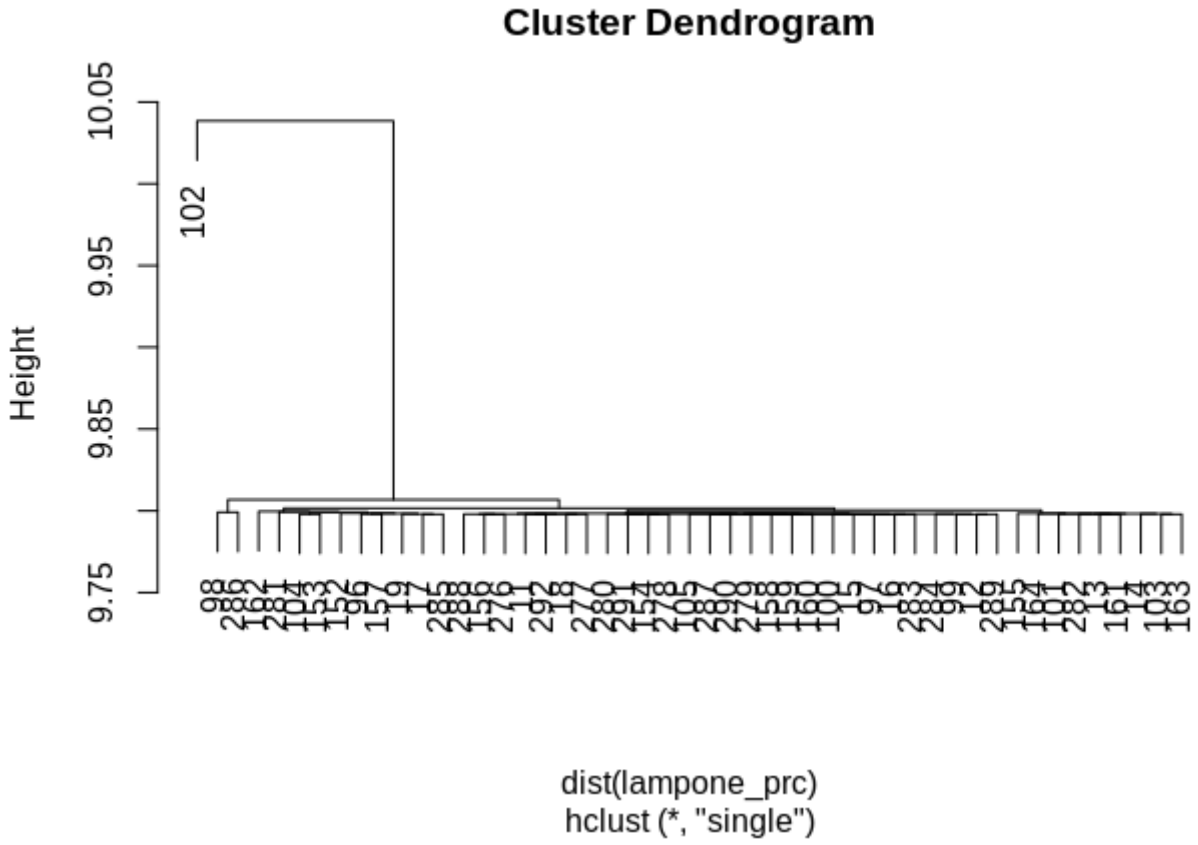


Figura 1.4: Dendrograma de hclust single sobre el dataset lampone transformado

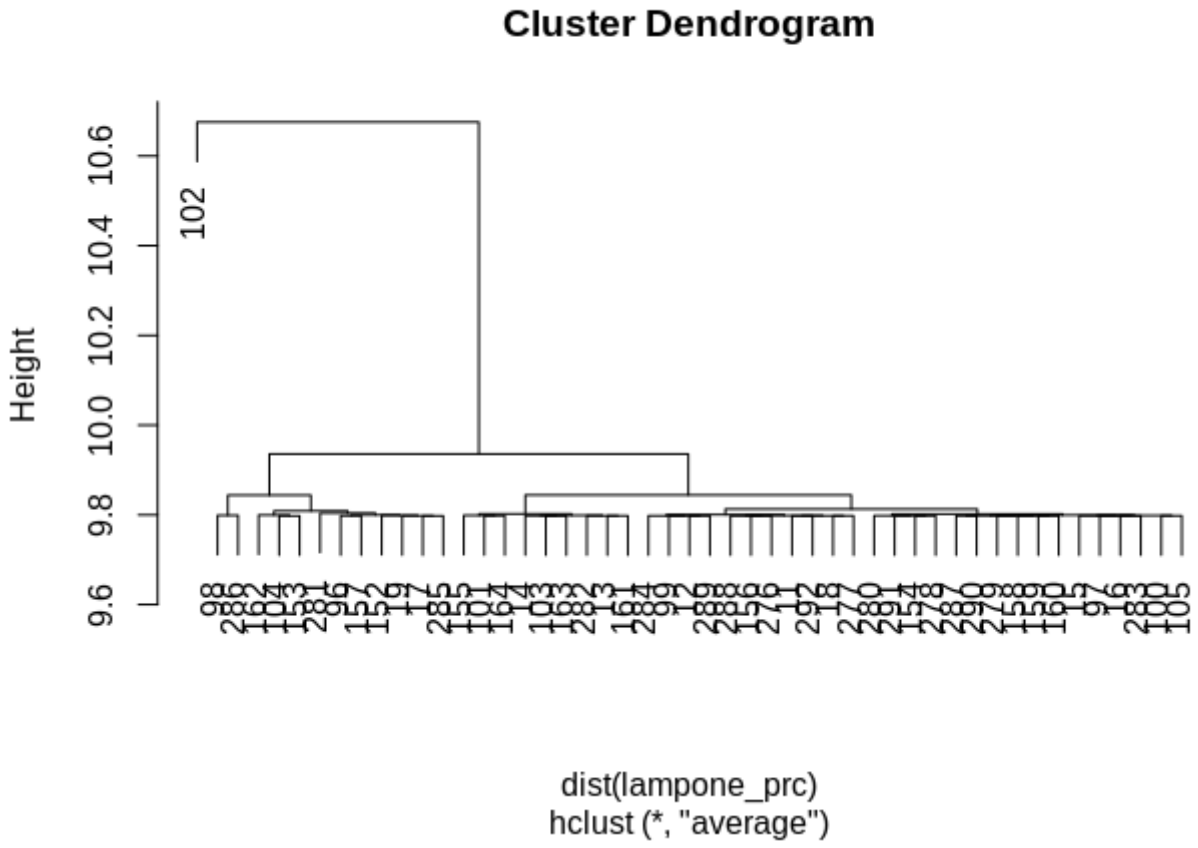


Figura 1.5: Dendrograma de hclust average sobre el dataset lampone transformado

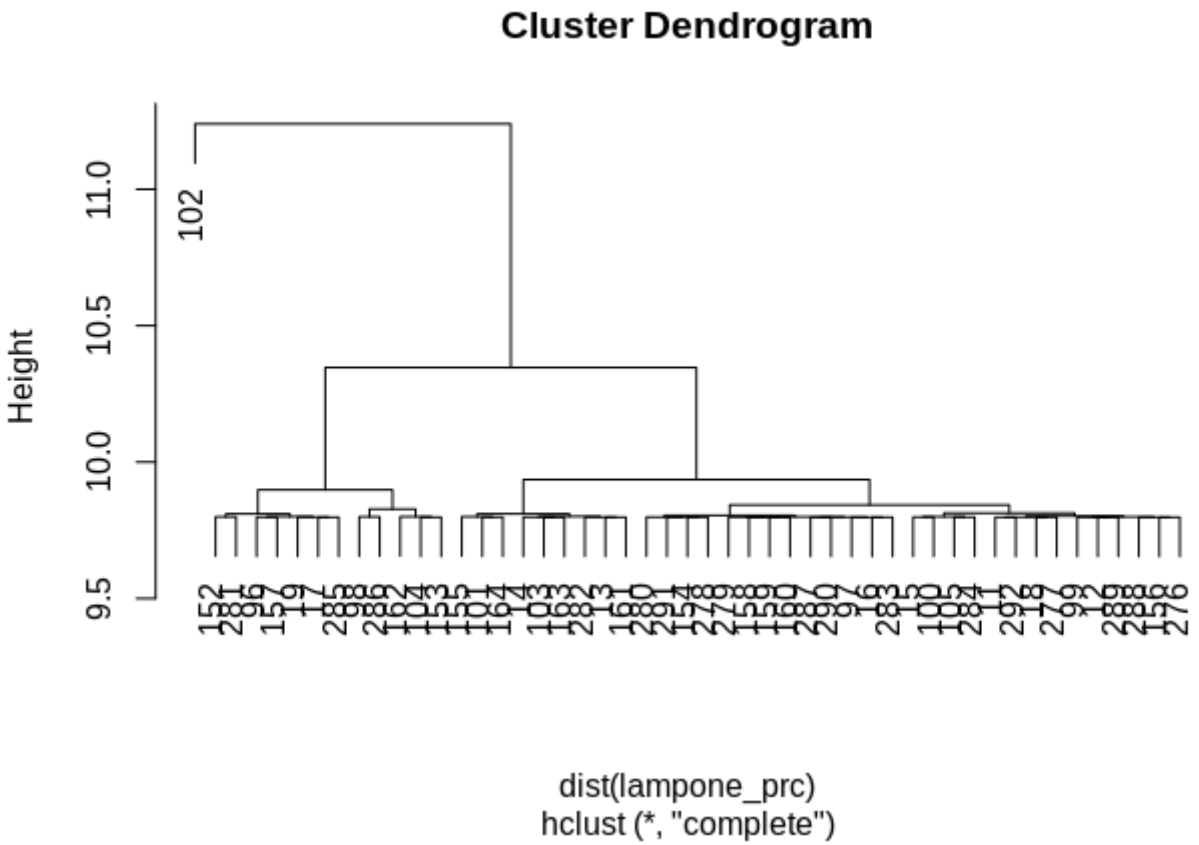


Figura 1.6: Dendrograma de hclust complete sobre el dataset lampone transformado

Al igual que el dataset de “crabs”, los métodos jerárquicos no obtendrán buenos resultados con nuestro dataset, (parcialmente) por los outliers de este.

Aunque todos los métodos parecen encontrar clusters más asociados al año que a la especie de frambuesa, la precisión de esta no llega a ser lo suficientemente relevante como para poder catalogar los resultados como “buenos”.

3) A continuación presentamos una tabla que indica la cantidad de clusters sugeridos, en la cual las filas representan el dataset usado y las columnas el método utilizado.

	GAP	Stability k-means	Stability hclust-single	Stability hclust-average	Stability hclust-complete
Gaussianas	4	3	3	3	5
Iris	2	2	2	3	2
Lampone	1	10	2	2	10

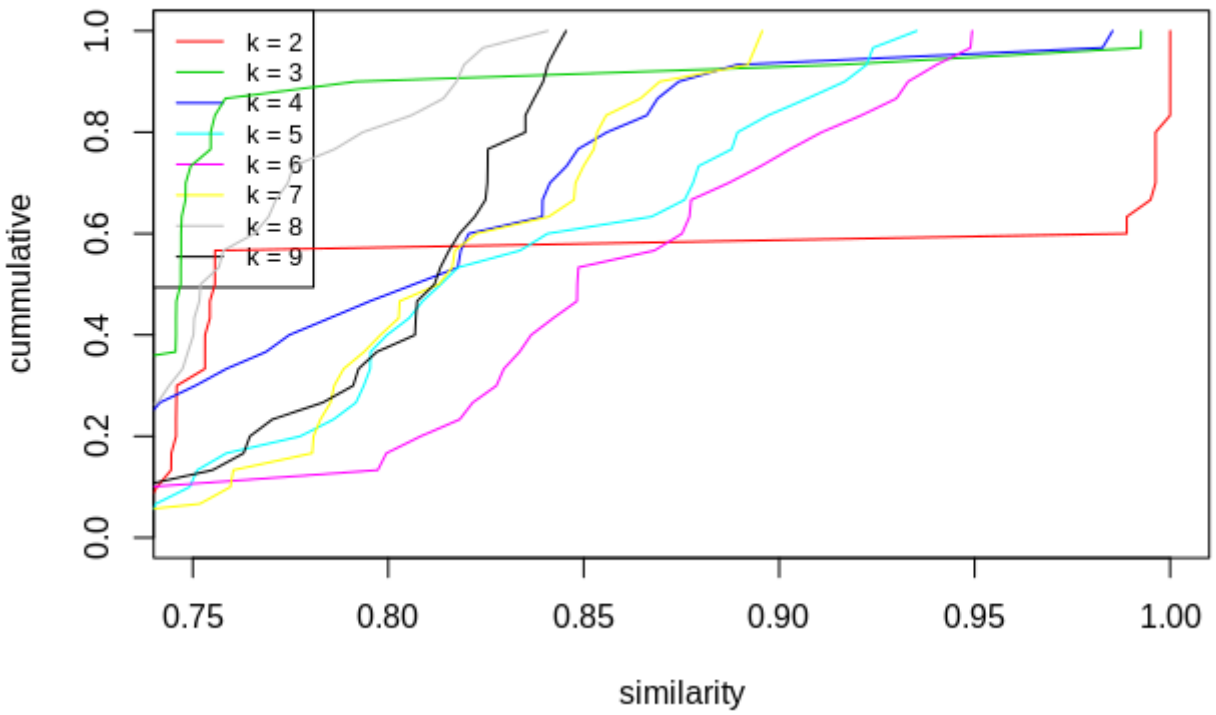


Figure 3.1: Stability Gaussianas kmeans

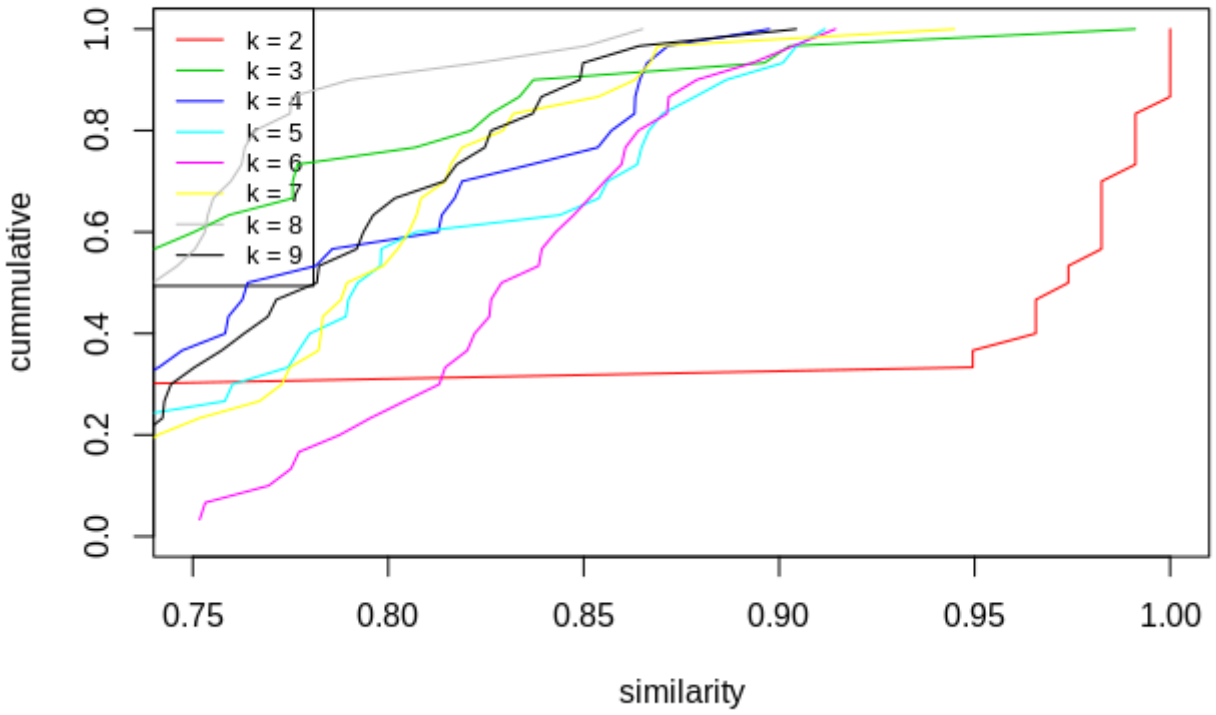


Figure 3.2: Stability Iris kmeans

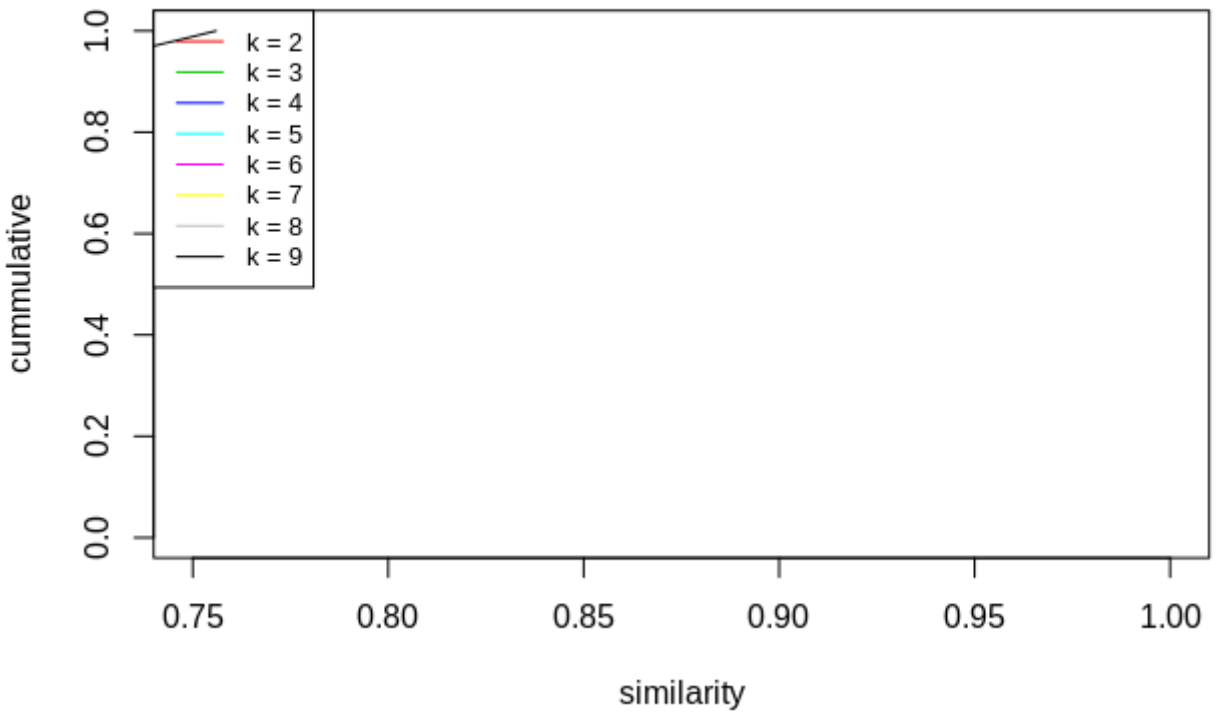


Figure 3.3: Stability Lampone kmeans

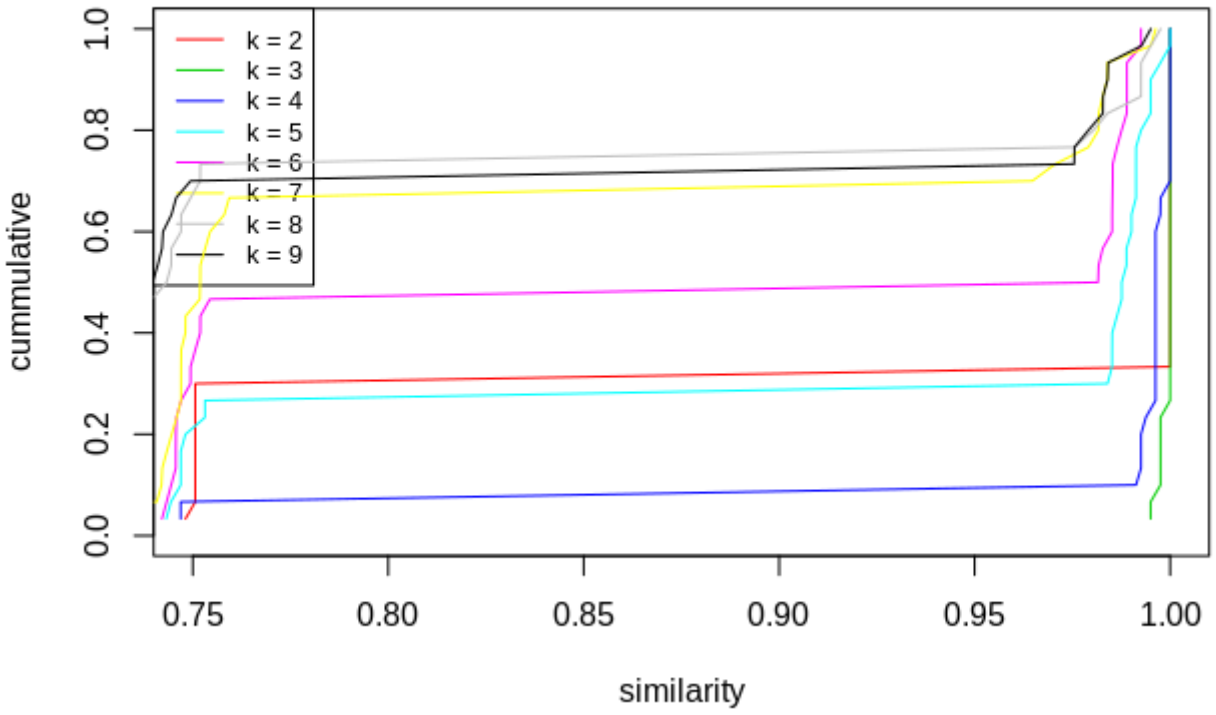


Figure 3.4: Stability Gaussianas hclust-single

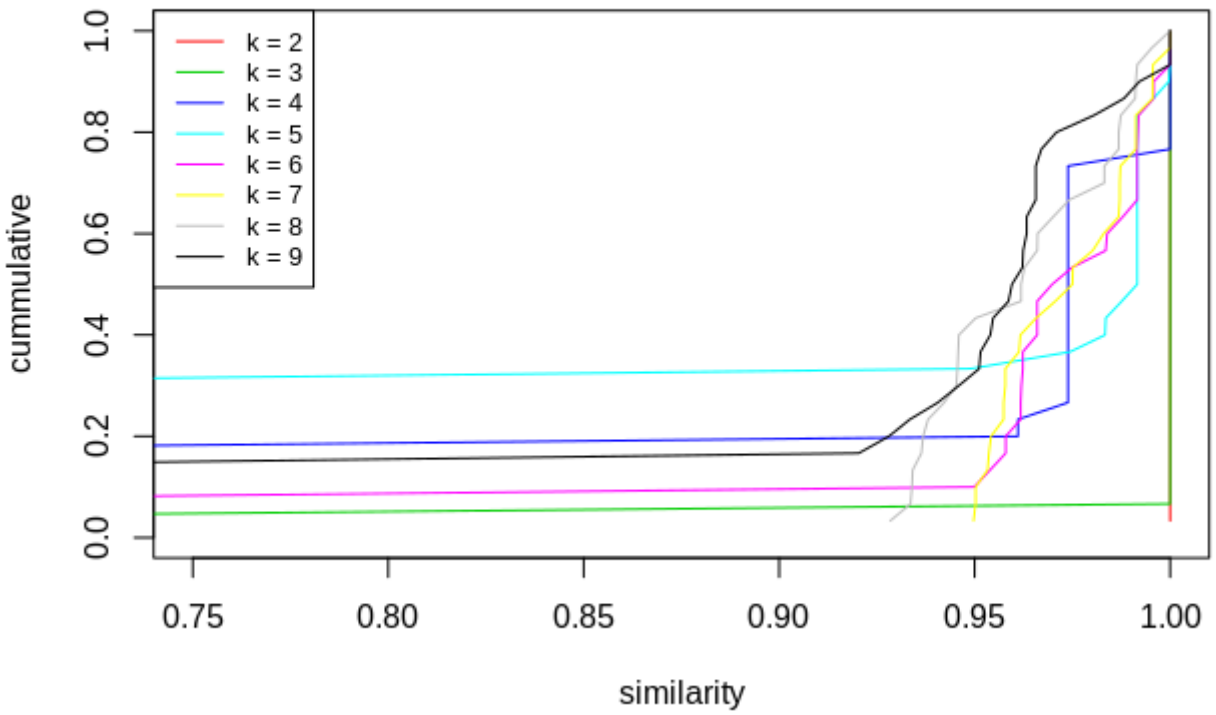


Figure 3.5: Stability Iris hclust single

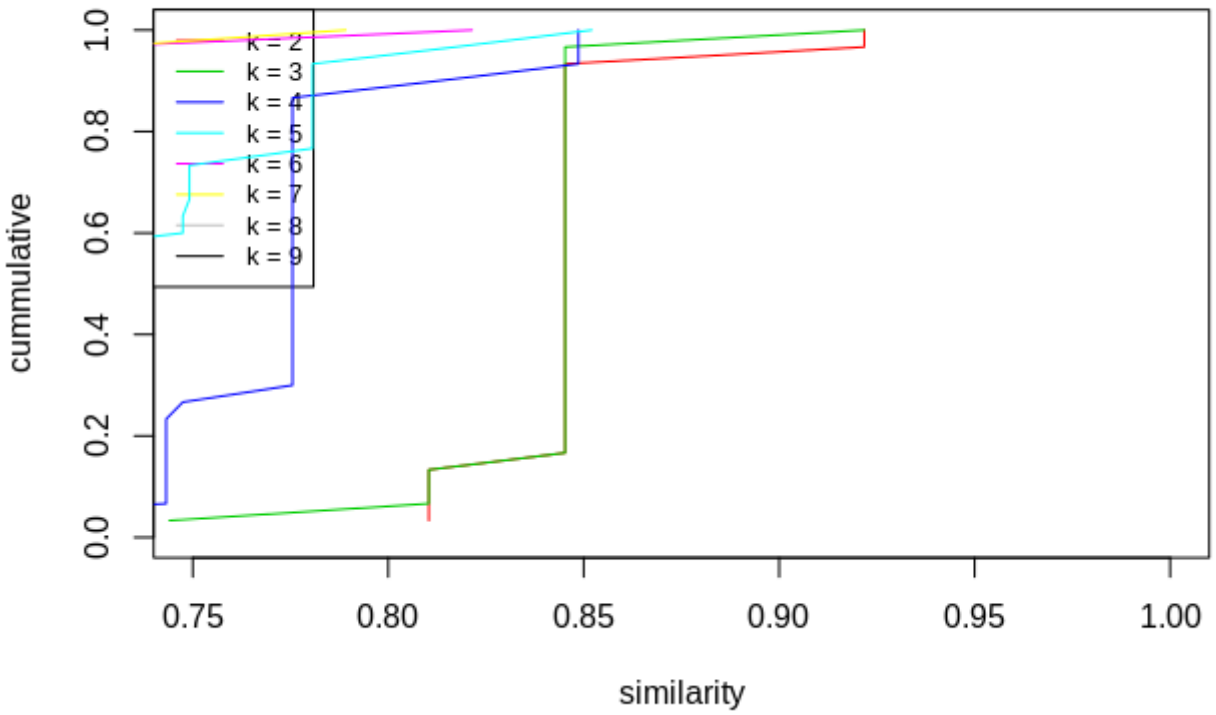


Figure 3.6: Stability Lampone hclust single

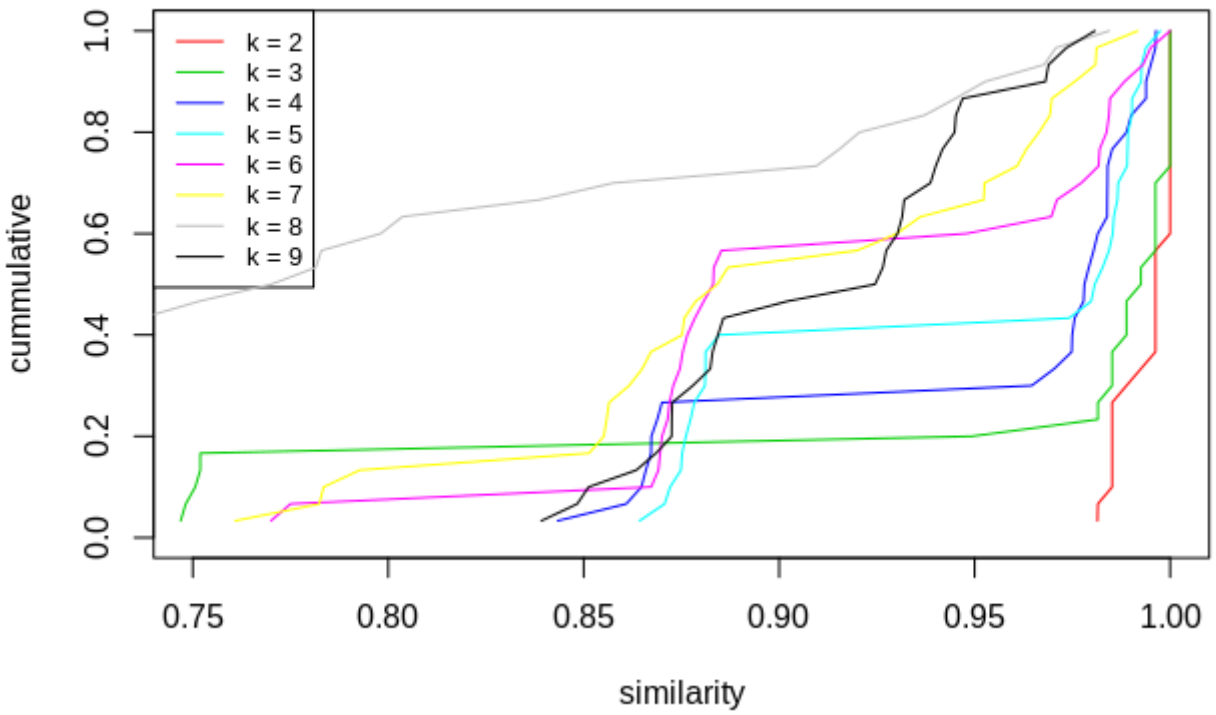


Figure 3.7: Stability Gaussians hclust average

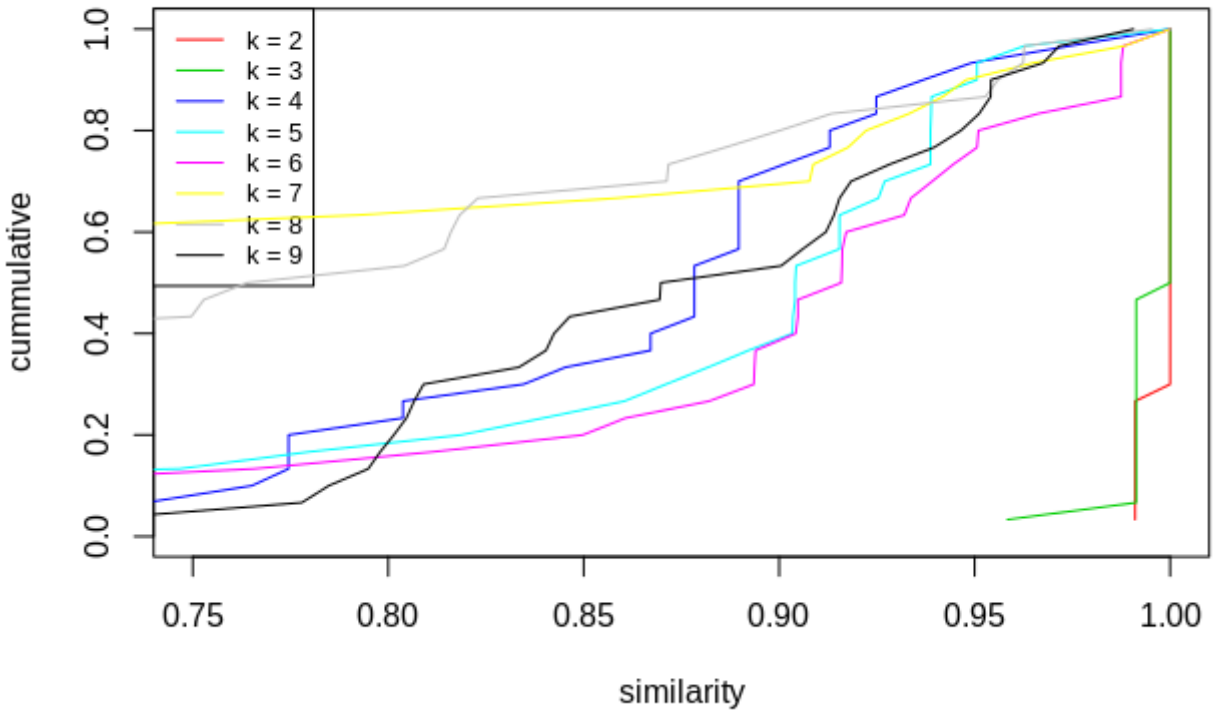


Figure 3.8: Stability Iris hclust average

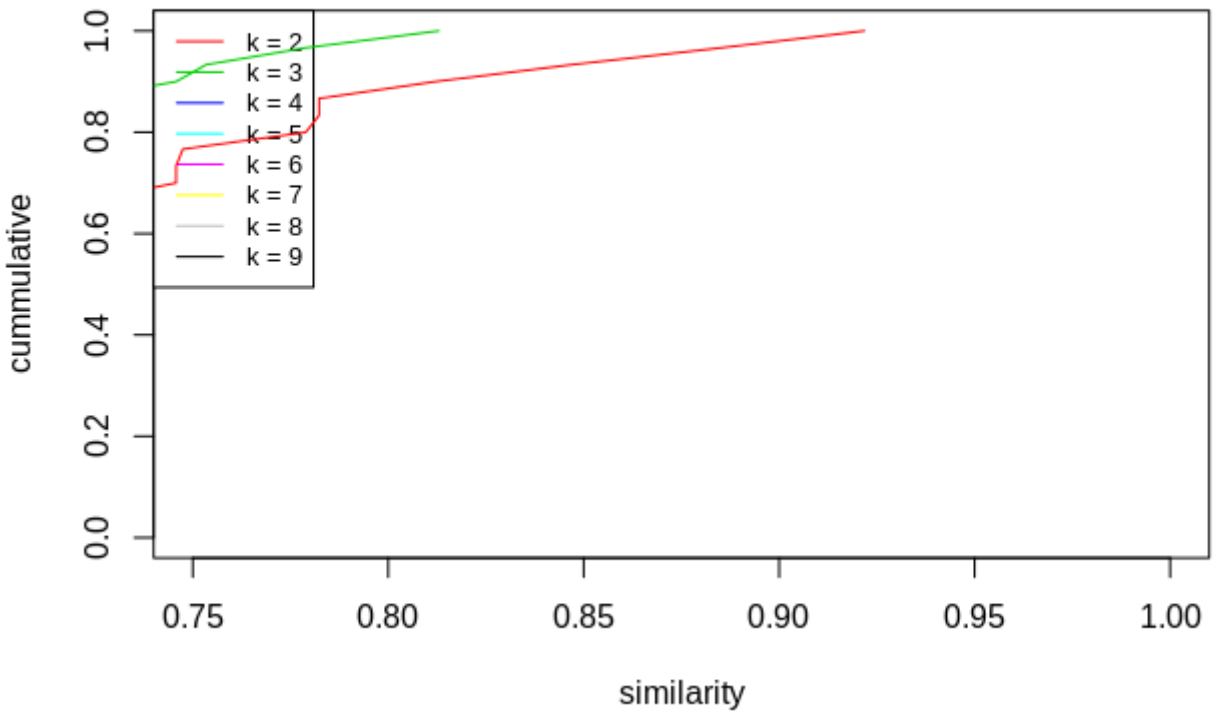


Figure 3.9: Stability Lampone hclust average

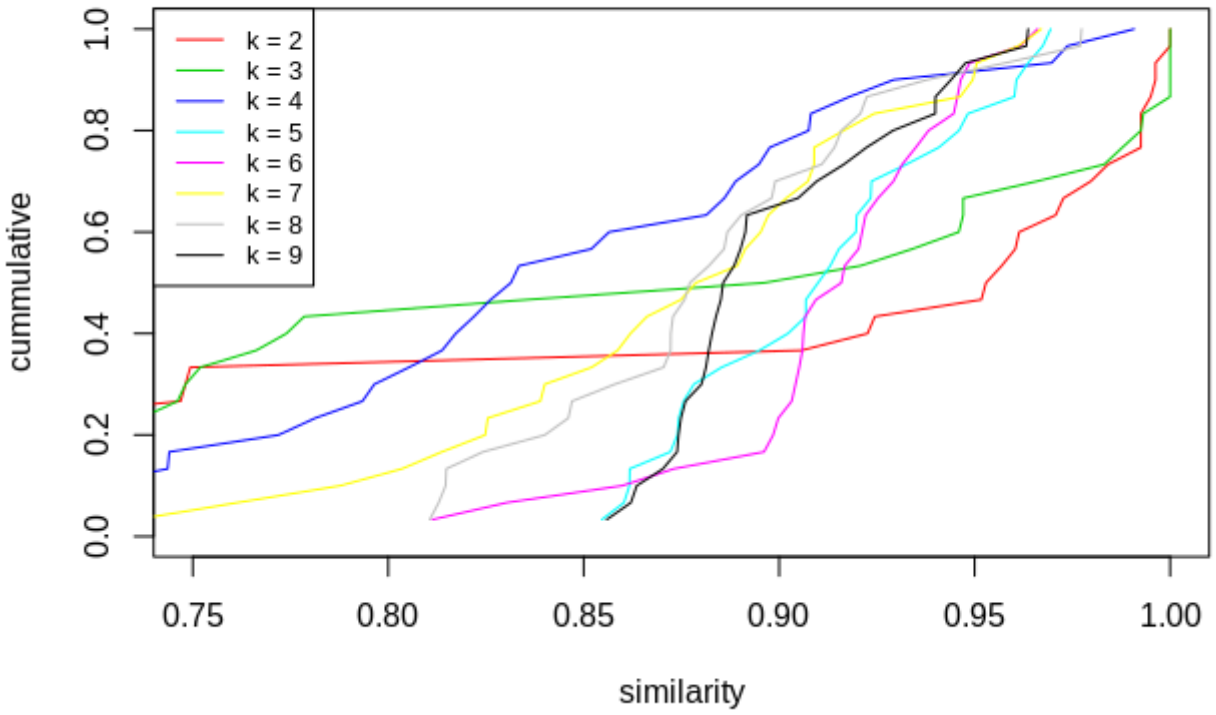


Figure 3.10: Stability Gaussianas hclust complete

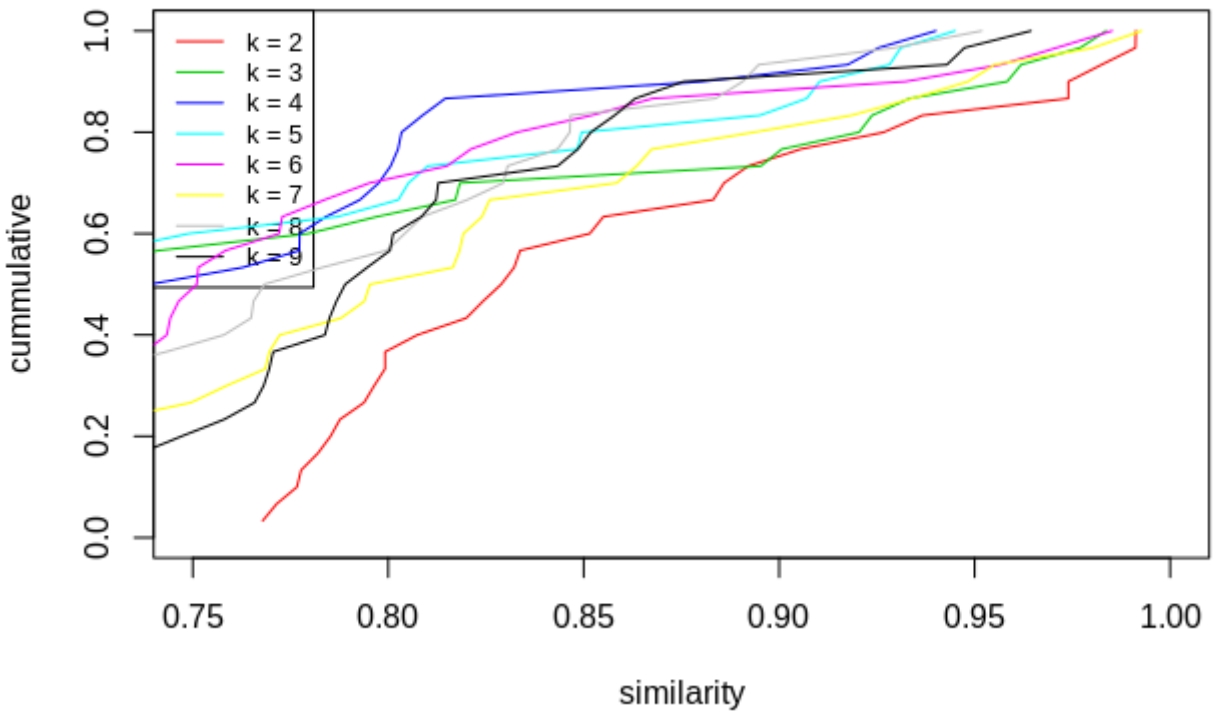


Figure 3.11: Stability Iris hclust complete

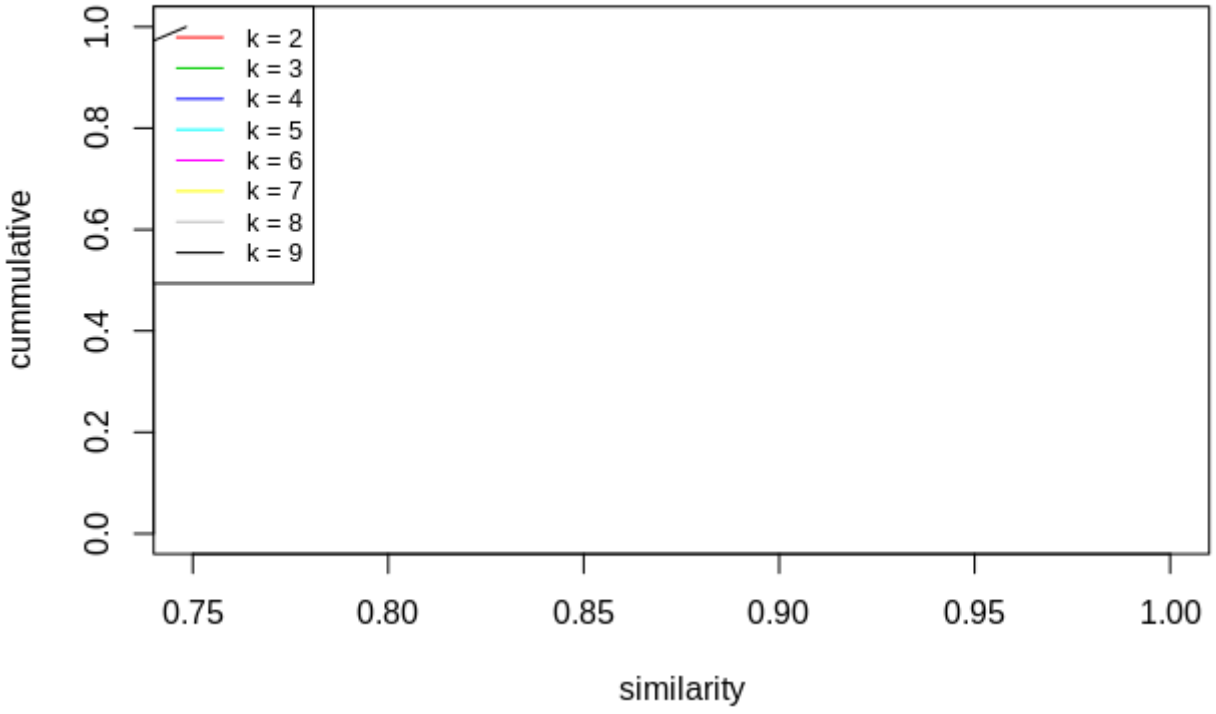


Figure 3.12: Stability Lampone hclust complete

Como las gaussianas son relativamente redondas tiene sentido que el método de Stability con kmeans de un resultado bastante cercano de clusters al real por la forma en la que funciona kmeans, buscando soluciones de puntos “apretados”.

Si miramos las variables de Iris (luego de pasar por ciertas transformaciones, entre ellas PCA) graficadas podemos entender por qué GAP y stability con k-means nos dan como resultado 2 clusters. Aunque pueden llegar a distinguirse 3 clusters al graficar iris, si vemos bien podemos ver como uno de los clusters está claramente más separado de los otros dos, por lo cual el “salto” al pasar de 1 cluster a 2 es más grande que el salto de 2 a 3.