# Taller 2: Análisis Multivariado

Análisis de conglomerados (Clusters)

Julián Camilo Riaño Moreno

viernes, junio 19, 2020

## Contents



## Actividad #2

> Busque una base de datos que contenga mediciones de varias variables cuanti- tativas sobre un conjunto de individuos de su interés.

### Especificaciones sobre la base de datos utilizada.

La base de datos utilizada para este ejercicio corresponde a la misma utilizada en los ejericicos 1 y 2 del curso en análisis multivariado en la especialización de estadística aplicada, durante el 1er periodo del 2020. Como se ha descrito en los ejercicios anteriores la base de datos corresponde a una modificación del *Data from the ECDC Surveillance Atlas*. De esta se obtuvo solo la información para dos especies de bacterias a saber: *Escherichia coli; Klebsiella pneumoniae*; y la información correspondiente a 30 países de 4 regiones de Europa (Sur, norte, este y oeste) durante los años 2000 al 2018. Se tienen 9 variables: `Bacteria`, `Year`, `Code_C`, `Country`, `Aminoglycosides`, `Carbapenems`, `Fluoroquinolones`, `cephalos_3er_gen`.

Para el análisis de conglomerados se decidió realizar una modificación para ajustar los datos a los requerimientos del análisis. En primera medida, para el análisis principal de esta actividad se seleccionó únicamente

la especie de bacteria que mayor cantidad de observaciones tenía *Escherichia coli*, seguidamente, se mantuvo como única variable categórica `Country` y luego se trasformaron las observaciones para cada país en la media de frecuencia de resistencia para cada uno de los antibióticos. De esta manera se obtuvo una base de datos de trabajo de 30 filas y 5 columnas (`Country`, `Aminoglycosides`, `Carbapenems`, `Fluoroquinolones`, `cephalos_3er_gen`. ).

Se realizó esto mismo para la otra especie estudiada en las actividades anteriores (*Escherichia coli*) y se realizó un análisis de conglomerados corto al final de este documento que reposa en el apartado de anexos.

## Primer pregunta

> Clasifíque los individuos usando cada uno de los métodos aglomerativos tratados en clase (método del vecino más cercano, método del vecino más lejano, unión mediante el promedio y el método de Ward) y usando la distancia euclidiana.

En esta primera parte del documento comentaré únicamente lo que corresponde al mapa de calor mostrado en la figura 1. el cual fue construido a través de la deterinación de las distancias eucledianas mediante la función `get_dist`del paquete `factoextra`. Lo que se refiere a los tipos de aglomerados y sus especificaciones se realizarán en el desarrollo del segundo punto.

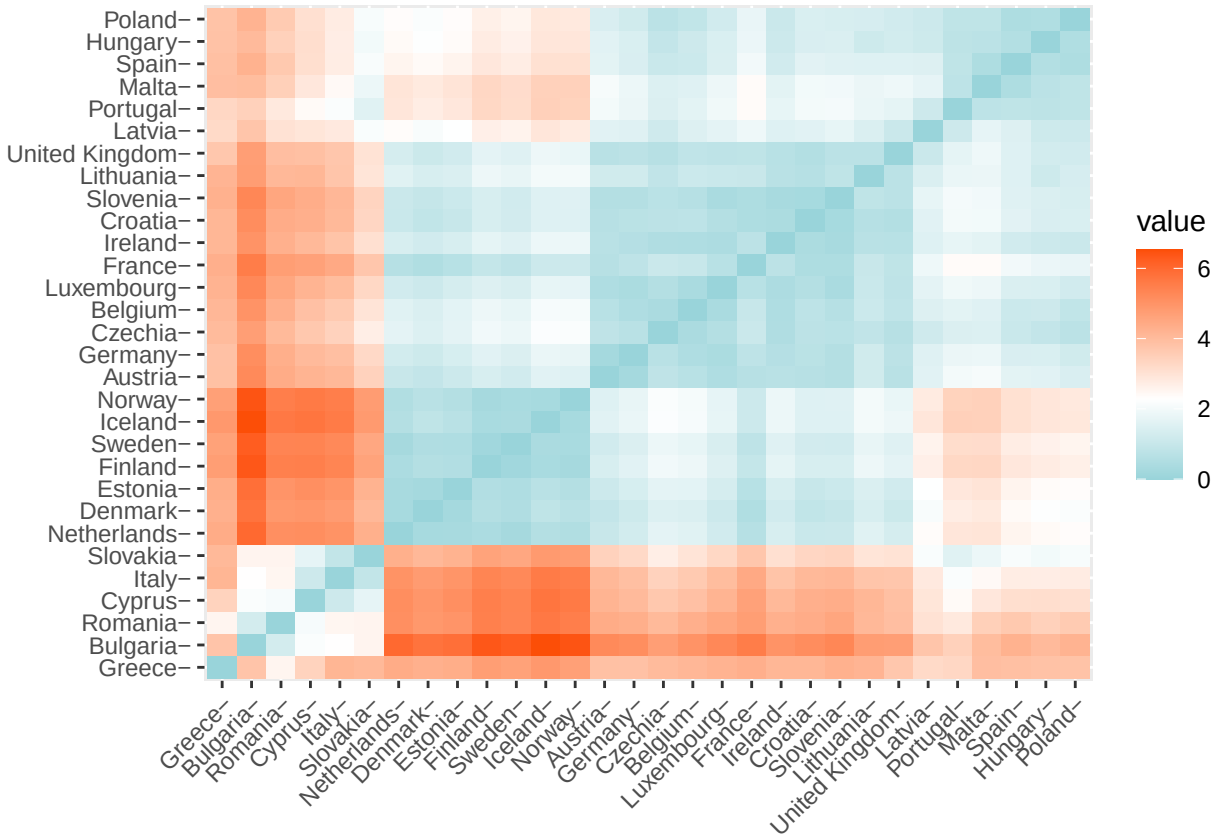


Figure 1: Mapa de calor de distancias euclidianas

La figura 1. muestra un mapa de calor que correlaciona las distancias eucledianas entre países construidas a partir de las medias de resistencia a los cuatro grupos de antibióticos definidos para estudio. En color rojo

se especifica las mayores distancias, es decir, la menor correlación entre los países respecto a la resistencia antibiótica y en azul se muestran las menores distancias entre países lo que sugiere mayor relación entre resistencia antibiótica entre los países.

Allí se puede ver claramente que el país *Bulgaria* presenta un patrón de gran distanciamiento con la mayor parte de los demás países excepto *Eslovaquia*, *Italia* y *Chipre*, con que muestra las menor distanciamiento. Ya en un análisis grupal se encuentra un gran conglomerado de países que presentan bajas distancias *Germany, Czechia, Belgium, Luxembourg, France, Ireland, Croatia, Slovenia, Lithuania, United Kingdom*, lo que sugiere que su comportamiento en cuanto a resistencia antibiótica es muy similar; al igual que el agrupamiento entre *Denmark, Estonia, Finland, Sweden, Iceland.*

Estos patrones de comportamiento llaman la atención a la luz del análisis de APC realizado anteiormente, dado que el primer grupo corresponde a países del sur de europa que como se vio allí son los que más aportan en la resistencia a los antibióticos en europa; En cambio los países de segundo y tercer grupo son los países que menos aportan a la resistencia antibiótica.

## Segunda Pregunta

> En cada uno de los casos del inciso anterior dibuje el dendrograma comente las diferencias entre cada uno de los resultados. Haga una caracterización de los grupos obtenidos.

**Métodos jerarquicos**

**Método de vecino más cercano**

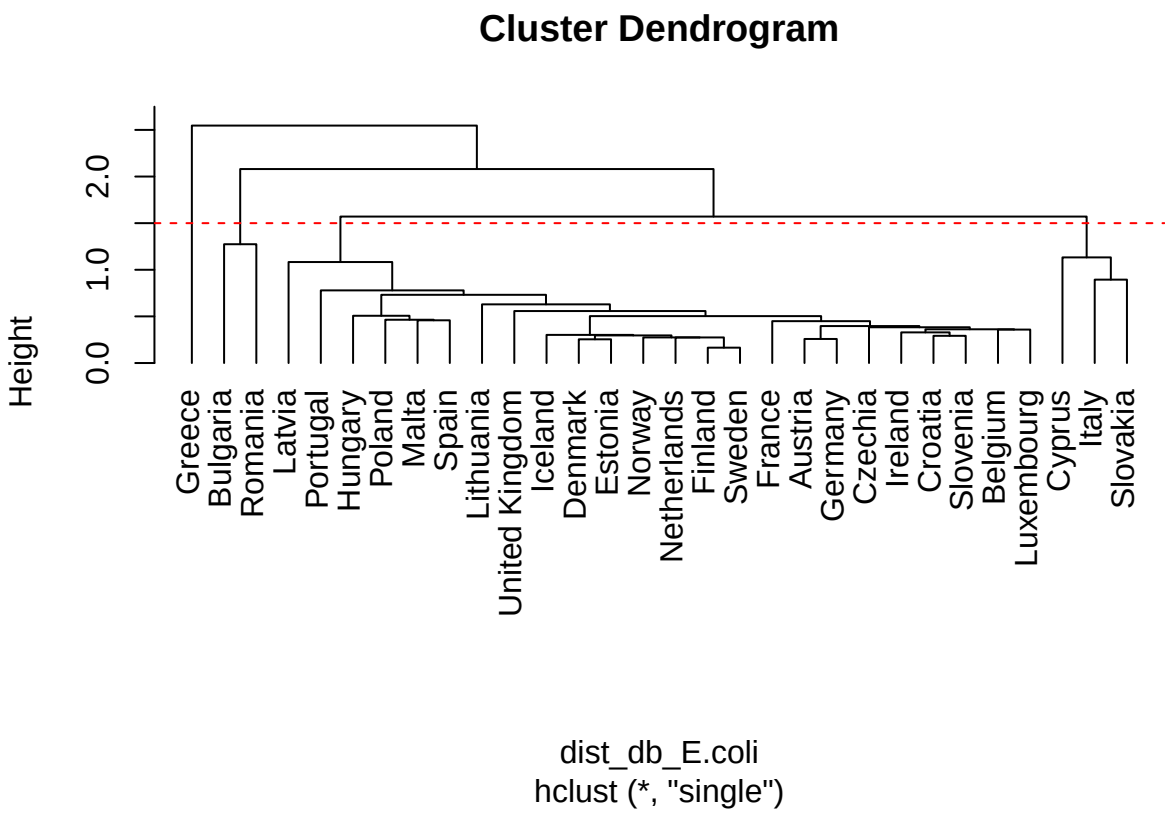


Figure 2: Dendograma utilizando el método *"vecino más cercano"*

En primer lugar se realizó un análisis a través del *metodo del vecino más cercano*, a partir de este método se construyo el dendograma mostrado en la figura 2. A partir, de esto se puede observar en primer lugar que el primer agrupamiento que se formó a través de las distancias eucledianas fue el que corresponde a `Finlandy` `Sweden`.

Al realizar un corte en el tamaño de distancia 1.5, este sistema de agrupamiento basado en las distancias más cercanas entre los elementos de los agrupamientos resultados, muestra que el comportamiento de la mayor parte de los países de centro y norte de europa respecto a la resistencia antibiótica a los 4 grupos de antibióticos es más cerca a la de cierto países de oriente de europa como `Slovakia` y países del sur de europa como `italia` y `Cyprus`(o Chipre), que a los países de `Bulgaria` y `Rumania` [1]. Llama especialmente la antención `Greece`, dado que corresponde a una rama independiente del arbol, lo que sugiere que su comportamiento en cuanto a resistencia antibiótica dista de la mayor parte de los países de europa.

[1]Esta rama es particularmente interesante por que los países `Bulgaria` y `Rumania` son cercanos geográficamente y culturalmente, esto podría llevar a suponer que su comportamiento en cuanto a la resistencia antibiótica puede estar dada por estas cercanías.

**Método de vecino más lejano.**

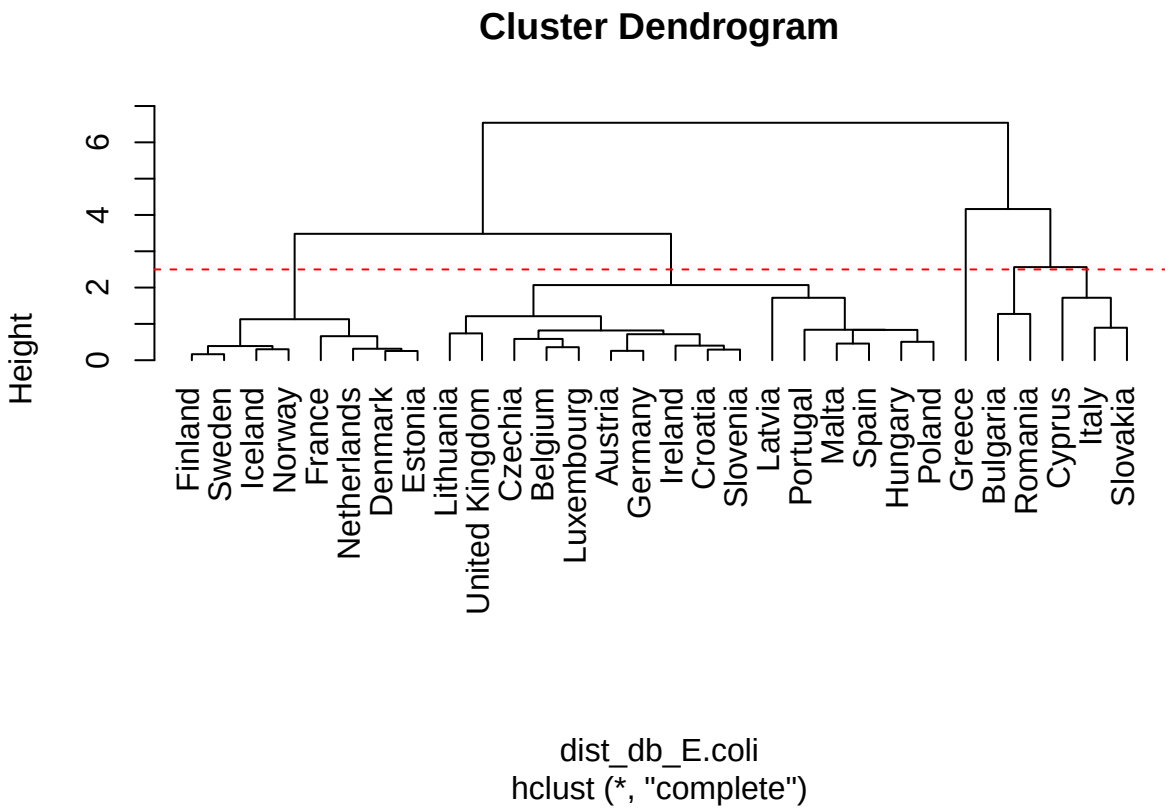


Figure 3: Dendograma utilizando el método *"Vecino más lejano"*

Con respecto al método del vecino más lejano, se contruye un dendograma (figura 3) en el cual se establecio de manera arbitaría pero que proporcionaría mayor significación a la altura de una distancia de 2.5. A través de este método que hace relaciones entre los aglomerados a través de las mayores distancias entre los elementos de cada aglomerado, se puede evidenciar un patrón similar al observado con el método de vecinos más cercanos antes descrito. `Slovakia`, `italia` y `Cyprus` conforman un aglomerado independiente, igualmente ocurre con`Bulgaria` y `Rumania` los cuales se ubican más cercanos a `Greece` que permanece aún lejano en comportamiento de resistencia a los antibióticos con el resto de los países europeos. El corte den 2.5 deja ver que el grupo central descrito anteriormente puede subdividirse en dos y esto muestra dos aglomeraciones una que corresponde particularmente a países del norte de europa (rama izquierda) y uno conformado principalmente por países del oeste de europa (rama central), Esto aunque sugiere que estos norte y occidente están cercanos en su comportamiento en cuanto resistencia antibiótica entre ellos pueden exisitir diferencias que vale la pena estudiar.

**Método de Ward**

Se elaboraron dos gráficas para mostrar los aglomerados a través del método de Ward (figura 4 y 5), estás gráficas únicamente son distintas en su manera de presentar los datos. En la figura 4 se estableción un punto de corte en la distancia entre conglomerados en 4. Con esto se pueden evidenciar 5 aglomerados, lo mismo se realizó en para la figura 5. donde se definió 5 grupos de aglomeramientos para estudiar.

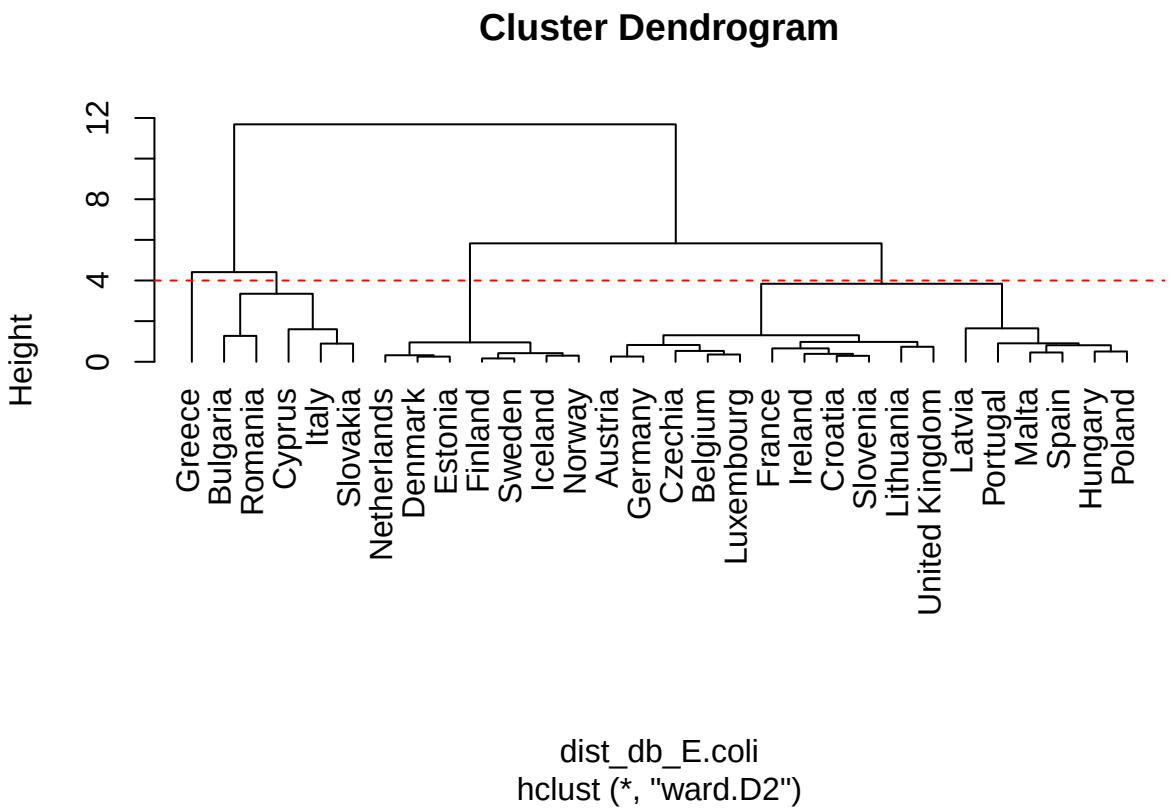


Figure 4: Dendograma utilizando el método *"Ward"*

El método de Ward los agrupamientos se realizan a través del efecto que exista en la varianza de los datos a medida que realiza uno u otro agrupamiento. Se definirá como agrupamiento al grupo de elementos que incremento lo más minimo el total de la varianza de todo el agrupamiento. Siguiendo este método se puede observar al igual que la figura 2 y 3, aglomeraciones de países del sur y oriente de europa (rama izquierda), el punto de corte en 4 y la aglomeración en 5, deja ver que `Greece`conforma un grupo externo al resto de los países lo que concuerda con un comportamiento particular en la resistencia antibiótica en este país.

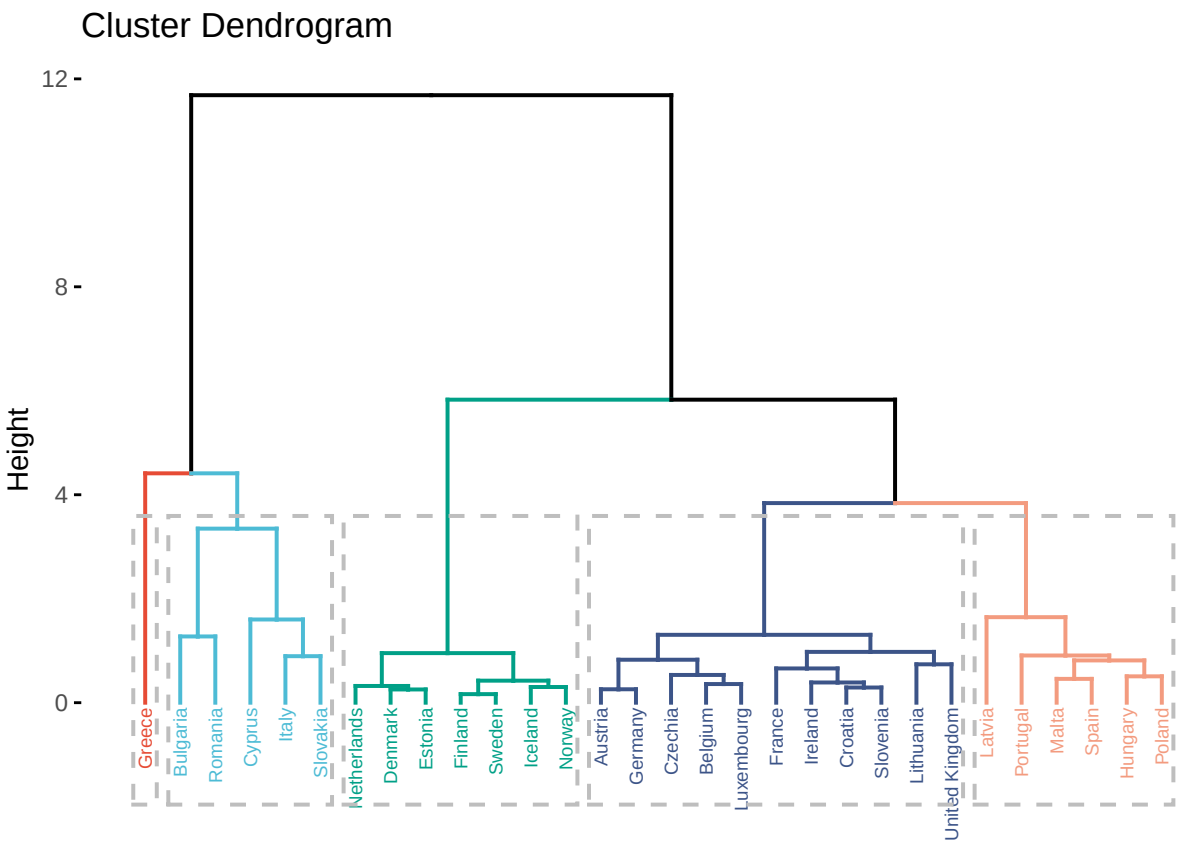


Figure 5: Dendograma utilizando el método *"Ward"* mejorado

**Método de promedio (*Average*)**

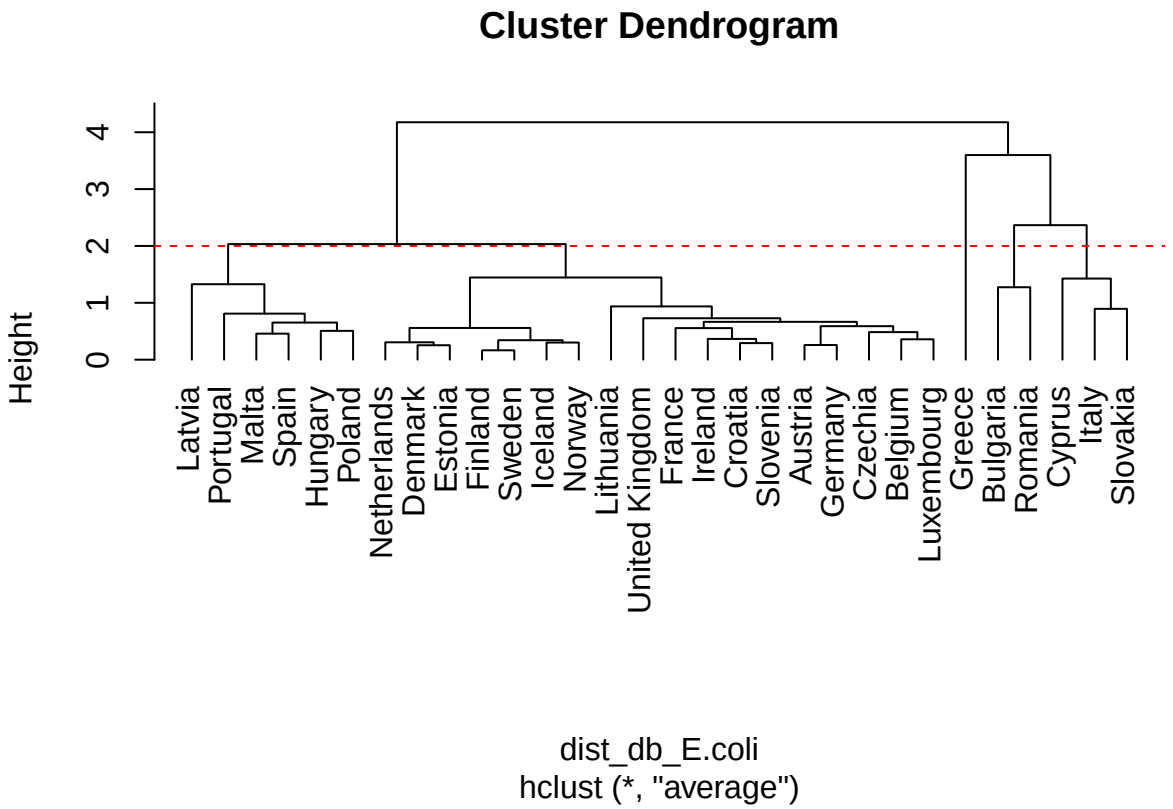


Figure 6: Dendograma utilizando el método *"Average"*

El método de promedio o *Average* se muestra en dos dendogramas (figuras 6 y 7). Este método realiza aglomeraciones a través de las medias entre las distancias de los elementos en cada uno de los aglomerados. Para este método se definió un punto de corte en 2 y 5 agrupaciones, esto permite ver mejor las aglomeraciones significativas para los países.

Inicialmente se puede observar, al igual que los métodos utilizados anteriormente un aglomerado de países de sur y este de europa. `Greece` nuevamente se encuentra formando una rama propia en el dendograma. Se ve además una aglomeración (rama izquierda) que muestra algunos países de oriente y sur de europa separada de los países de centro y norte europa, lo que sugiere su comportamiento similar pero con algunas distinciones que deberían ser evaluadas.

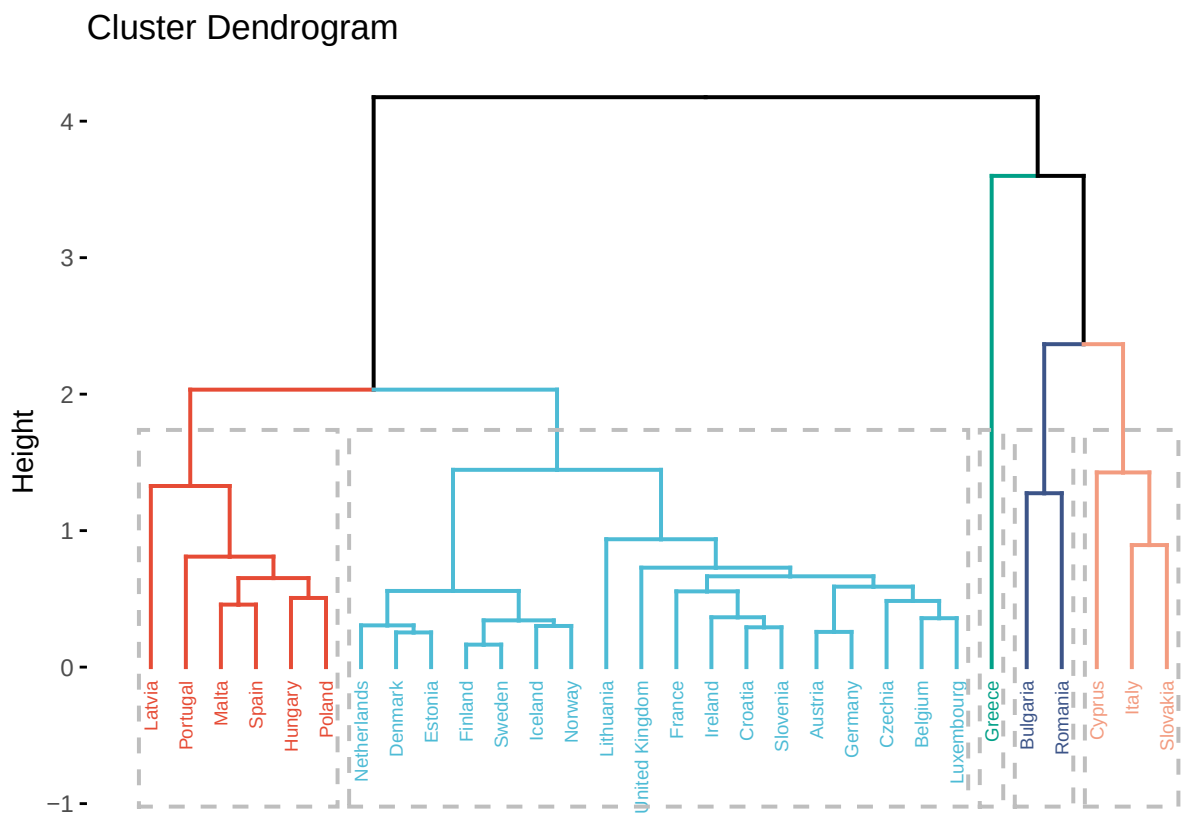


Figure 7: Dendograma utilizando el método *"Average"* mejorado

## Tercera pregunta

> Usando el método de $K$-medias realice el análisis de agrupación de los individuos y haga un gráfico que muestre los grupos diferenciados por colores. Interprete los resultados.

### Método no jerarquico ($K-medias$)

Finalmente se procede a realizar una análisis de aglomerados a tráver de un método no jerarquico conocido como K-medias. Para este método se define un centroide al azar, luego se toma cada observación del aglomerado y se revisa con relación o distancia al centroide (vector de medias), el objetivo es minimizar esta distancia. Esto se realiza con un número definido de iteraciones (para este ejercicio se realizaron 25) y el aglomerado se determina a partir de los aglomerados que menor distancia tengan de un centroide determinado.

Previamente se decide realizar un análisis del criterio de *Mojena* para definir el número de agrupamientos que se requiere para este análisis. Para esto se define una función llamanda `mojena`y se gráfica a través de la figura 8. en esta imagen se puede observar que el criterio de *Mojena* responde al punto dónde la distribución de la sumatotal deja de ser significativa esto es el punto de caida dónde los datos no ofrecen mayor información. Al aplicar este criterio se obtiene que el número de aglomerados más adecuado para este grupo de datos es 4.

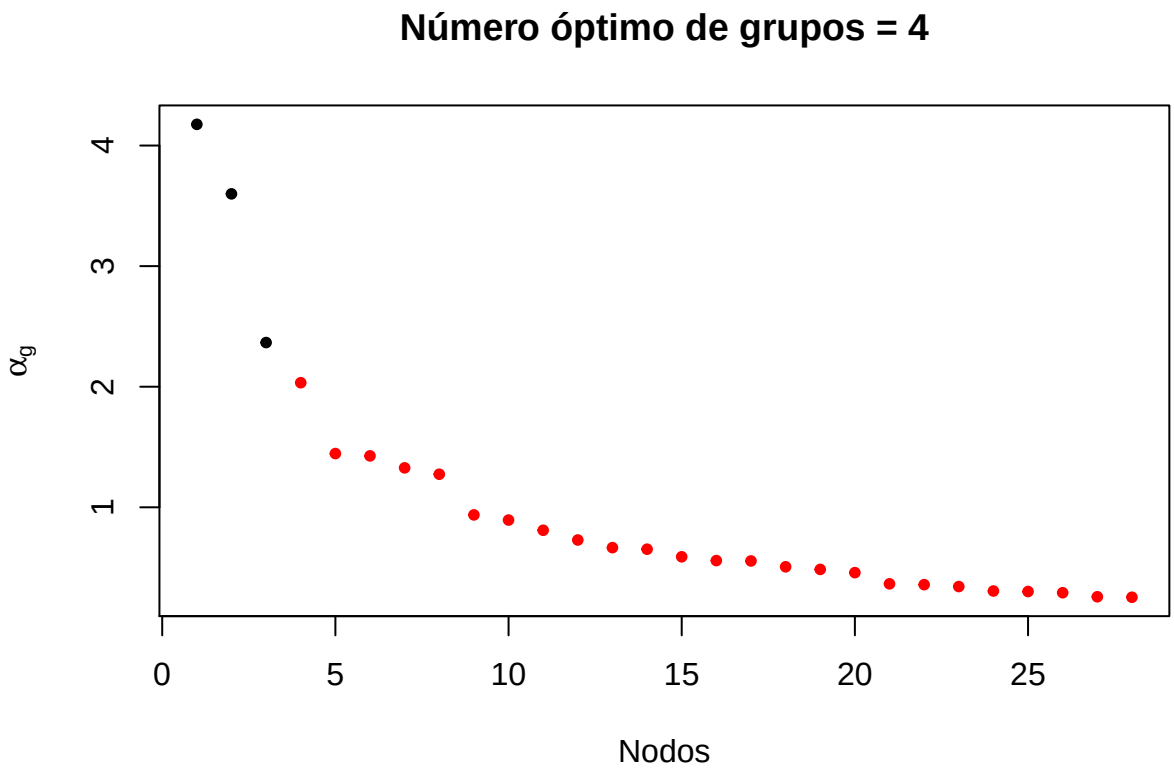


Figure 8: Gráfica de Mojena para identificación de número óptimo de conglomerados

Se comprueba esto a través del método del *codo* el cual se muestra en la figura 9. Este método define que el número de aglomerados es aquellos que estén dentro de la suma de los cuadrados, en el punto de inflexión dónde esta suma no aporte más información. Al aplicar este método se puede observar nuevamente que el número de aglomerados más adecuado para este conjunto de datos es 4.

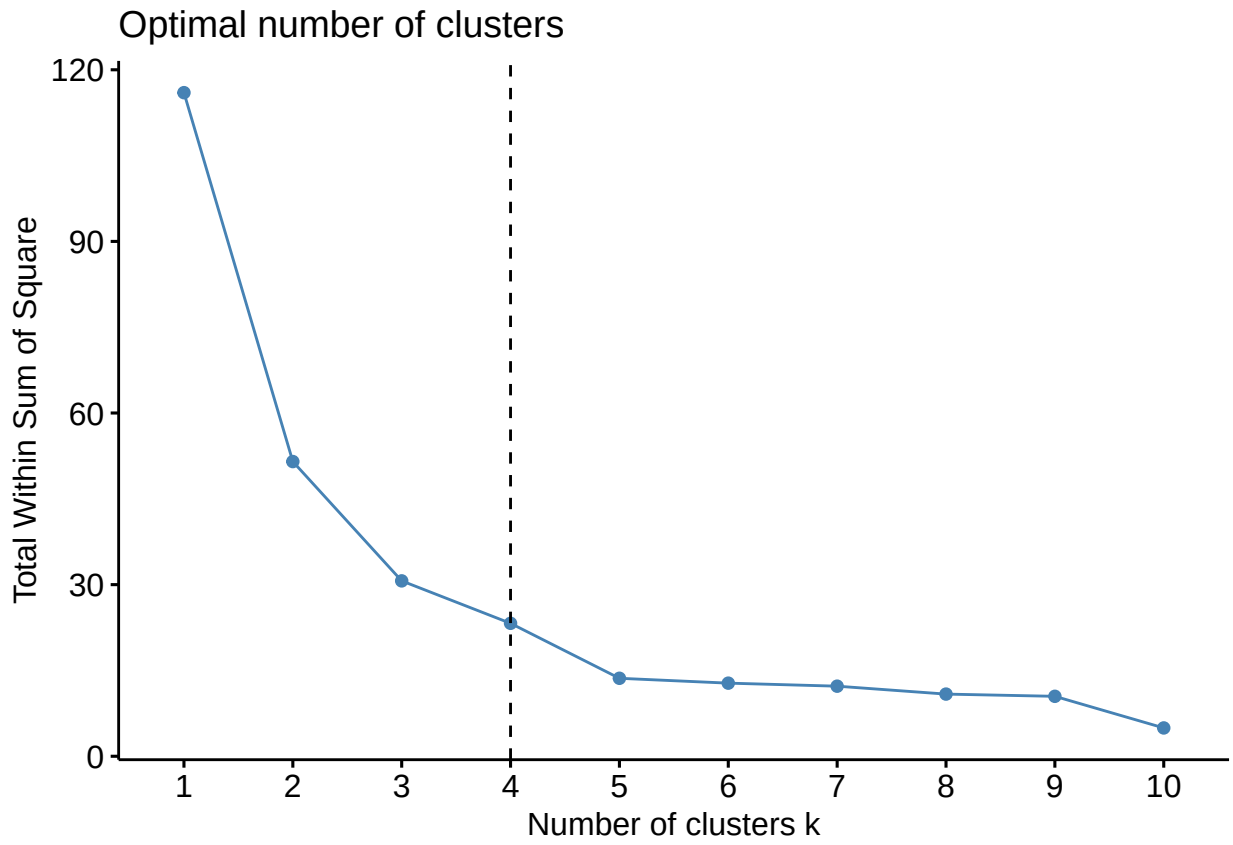


Figure 9: Gráfica para definición de número de conglomerados para $K - medias$ por método WSS (within-cluster sum of square)

Definido el número de aglomerados, se procede a gráficar estos a través del método de K-medias (figura 10). Esta figura llama la atención encuanto define los 4 aglomerados, donde uno de estos está conformada por una única observación, esta corresponde a `Greece` esto quiere decir lo que se ha visto anteriomente que este país presenta una problemática en cuanto a resistencia antibiótica muy particular que no se relaciona con los comportamientos de los demás países. Además se puede observar que el conglomerado 3 (gris), agrupa principalmente países de europa del sur y oriente y su aporte lo realizan en la primera componente principal, esto indica que estos países son los que más aportan a la resistencia atibiótica en Europa (junto a Grecia) en especial a antibióticos como las `Aminoglucosides` y `Cephalo_3er_gen`.

De igual manera, es interesante el análisis de los agrupamientos 2 y 4. Claramente se observa que los países que corresponde al grupo 4, son países del norte de europa y son países cuya contribución es poca para la problemática de resistencia antibiótica. Los países del grupo 2, principalmente de países de occidentales también tiene un escaso impacto en la resistencia antibiótica, incluso se puede observar en este grupo algunos páises de sur y oriente de europa. Sin embargo, se evidencia que son países que son geográficamente más cercanos a países occidentales, esto puede sugerir que los patrónes de resistencia en Europa presentan un patrón geográfico partícular y que las acciones frente a la resistencia a los atibióticos pueden ser más efectivas en países de occidentales y del norte de europa.

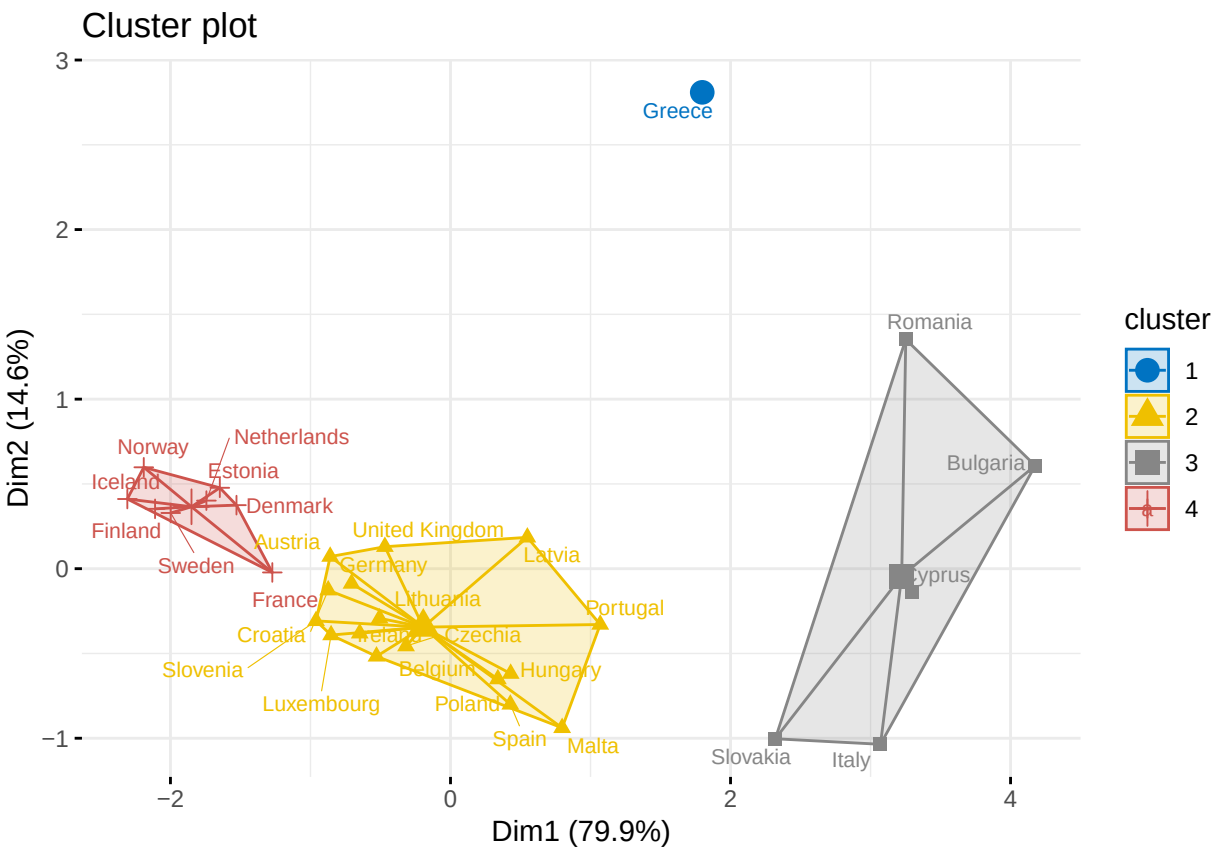


Figure 10: Gráfica de conglomerados por método de K-medias

## Anexo (análisis de conglomerados ($k$-medias) para Klebsiella Pneumoniae)

Para ampliar el análisis previo se decide realizar un estudio del comportamiento entre países de la especie bacteriana *Klebsiella pneumoniae*, para evaluar si el comportamiento de resistencia los grupos de antibióticos estudiados se asemeja al visto al ya descrito en el análisis anterior de la *Escherichia coli*,

La figura 11. permite apreciar para la *Klebsiella pneumoniae* algunas similitudes con el mapa de calor de distancias descrito antes para la otra bacteria. Sin embargo, en acá se ve que la distancias de resistencia para Klebsiella en `Greece` dista de los demás países en Europa. Esto concuerda con lo visto en el caso anterior.

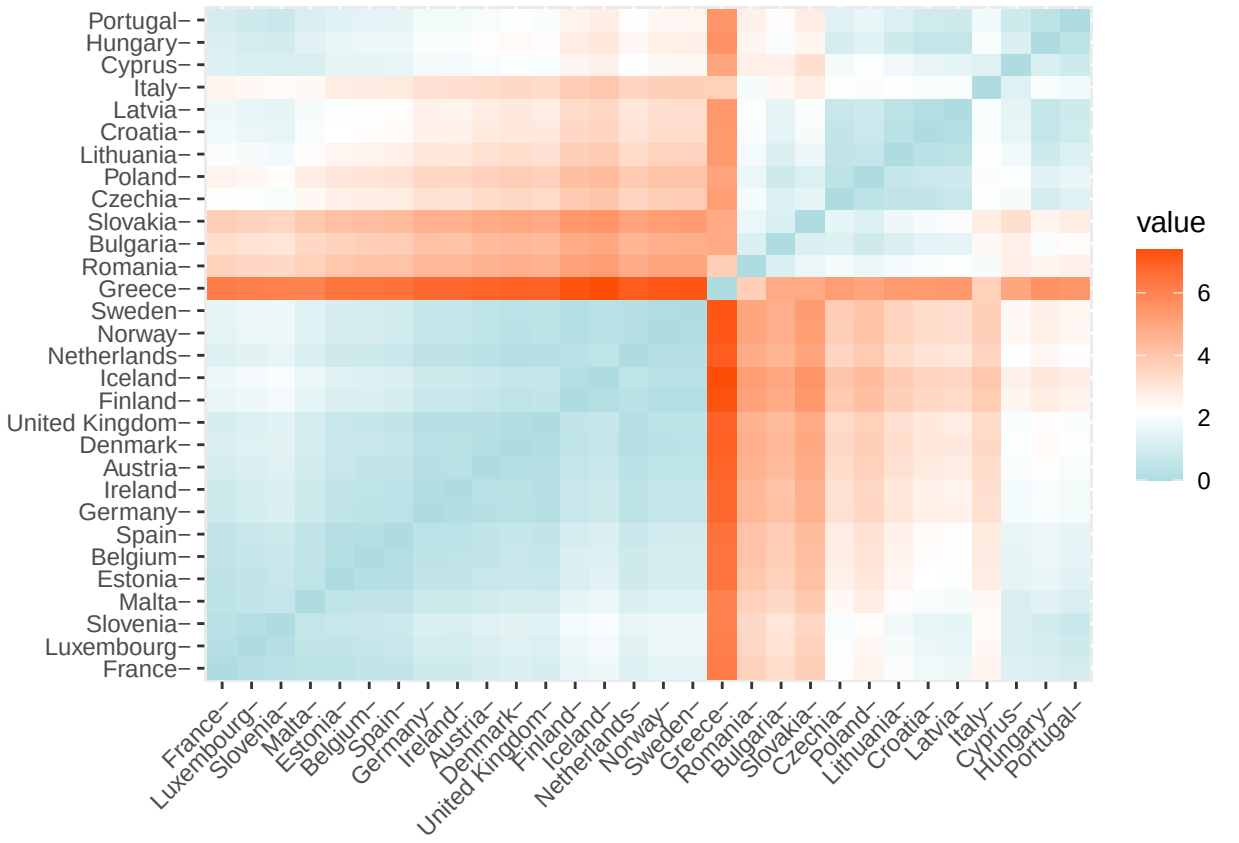


Figure 11: Matriz de distancias euclidianas para Klebsiella Pneumoniae

Se realiza un dendograma a través del método *Average* encontrandose nuevamente una fuerte distinción de aglomerados entre países de sur y oriente de europa con relación a los de norte y occidente, y nuevamente `Greece` se encuentra en una rama única lo que sugiere su comportamiento único respecto a la resistencia antibiótica de la Klebsiella.

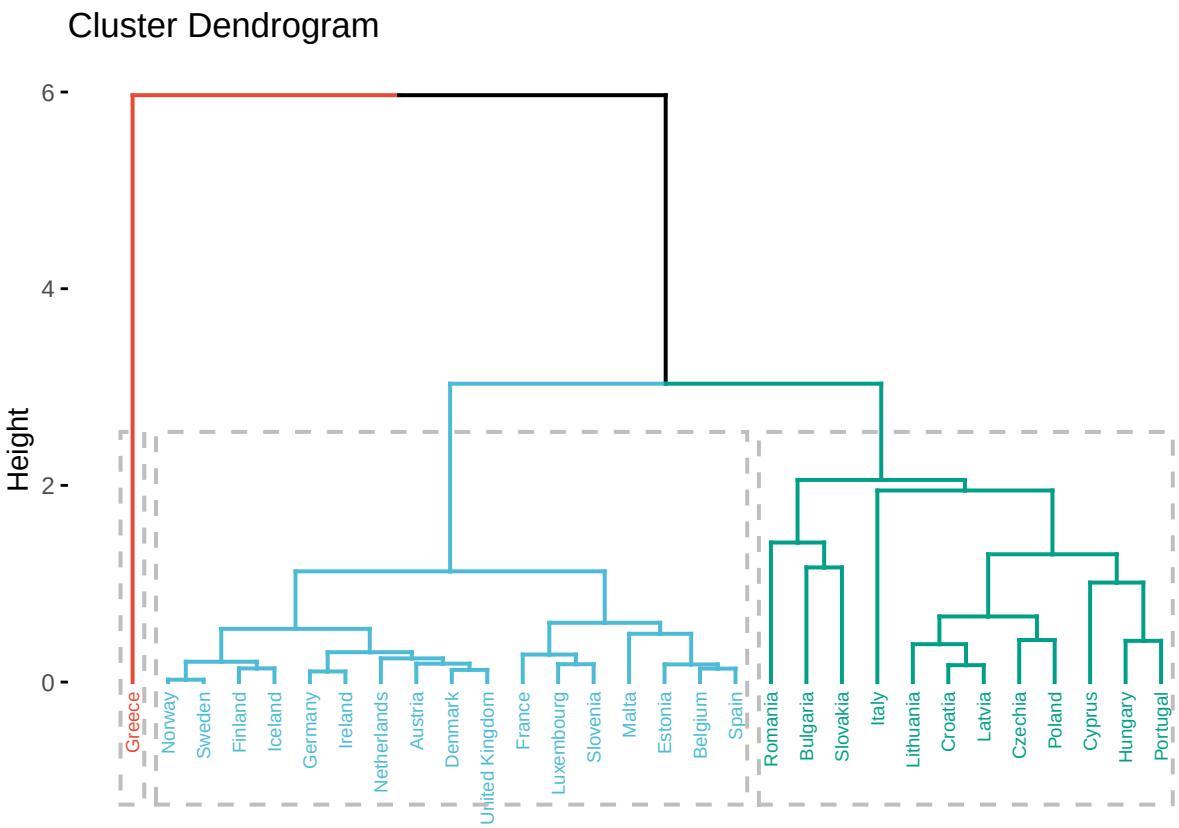


Figure 12: Dendograma utilizando el método *"Average"* para Klebsiella Pneumoniae

Se decide aplicar el criterio de *Mojena* (figura 13.)para definir el número de aglomerados para definir en el método de $K - medias$, obteniendose 3 aglomerados, lo que es comprobado por el método del *codo*

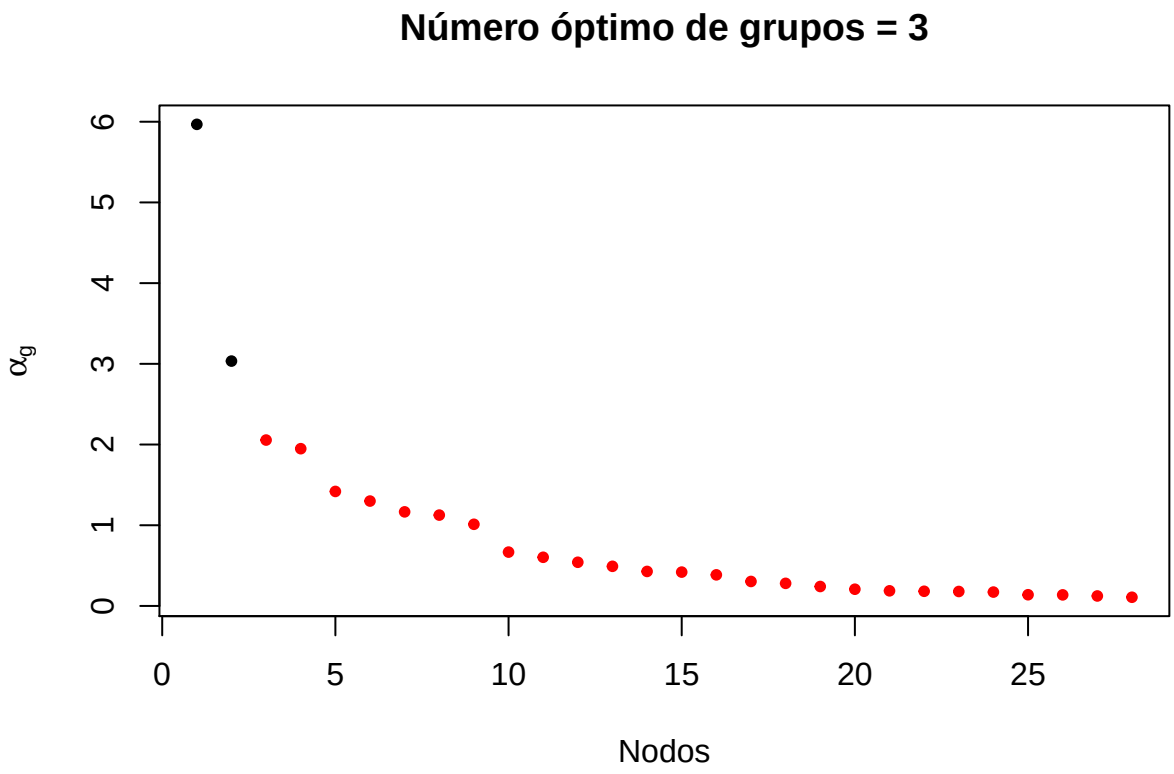


Figure 13: Gráfica de Mojena para identificación de número óptimo de conglomerados para Klebsiella Pneumoniae

A partir de la definición de estos tres aglomerados se decide gráficar el método de $K - medias$ con 25 iteraciones (figura 14.). Nuevamente se observa un grupo propio para `Greece` es el mayor contribuidor para la componente principal 1; esto indica que es el país que más conribuye a la resistencia antibiótica de la *Klebsiella pneumoniae* para Fluoroquinolonas y cefalosporinas de 3er generación. Además se evidenica que siguiente grupo que contribuye es el 2 (verde) el cual está conformado por países de sur y oriente de europa y nuevamente los países de norte y occidente son los que menos contribuyen a esta problemática.

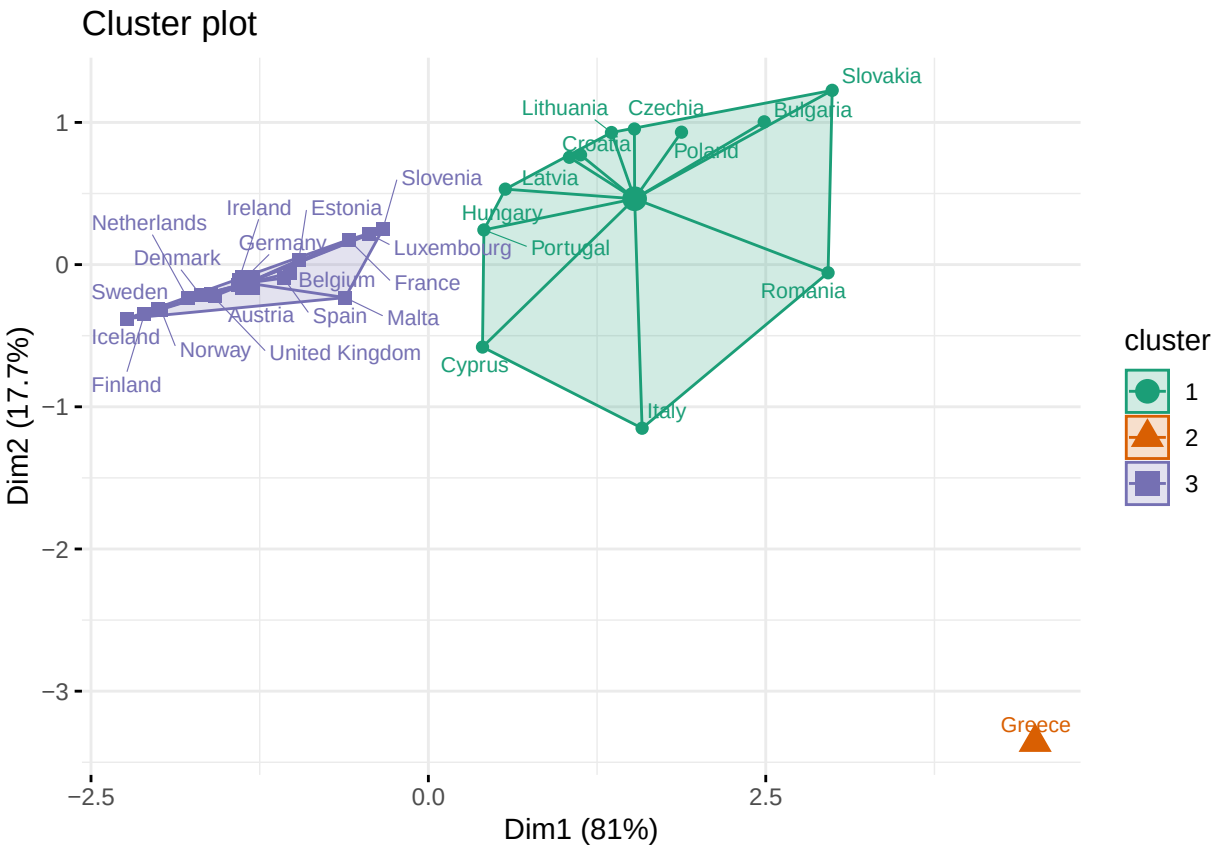


Figure 14: Gráfica de conglomerados por método de K-medias para Klebsiella Pneumoniae

## Conclusion y perspectivas

- Los países del sur y oriente de europa son los que más contribuyen a la resistencia antibiótica tanto de *Escherichia coli* como de la *Klebsiella pneumoniae*. Se hizo una pequeña revisión de la literatura y se encontraron resultados similares para otros agentes infecciosos; por ejemplo, esta región es la que más resistencia antibiótica en diferentes especies de *Acinetobacter*. Además recientemente se encontró que los países de estas regiones son las que más tienden a utilizar antibióticos de última línea. de manejo para agentes infecciososo como *Pseudomonas aeruginosa* y diferentes especiedes de bacterias del genero *Enterobacteriaceae*. Incluso la OMS (organización mundial de la Salud).ha previsto la gran necesidad de aunar esfuerzos para resolver la situación de resistenica antibiótica en estos páises

- Dentro de los países sur-orientales de europa el que tiene un comportamiento más particular es Grecia. Parece que es el máximo contribuidor de resistencia antibiótica en las dos bacterias estudiadas, esto indica que sus comportamientos para el manejo antibiótico son distintas a las vistas en otros países. Esto se ha intentado explicar por un abuso del uso de antibióticos en diferentes escenarios clínicos en este país . No es la primera vez que se evidencia esto, ya múltiples grupos han encontrado resultados similares. Sin embargo, las razones precisas de este efecto no son del todo claras y se requieren de más estudios de indole social, cultura y clínico par establecer este efecto.

- En el análisis llevado acabo en esta actividad se evidencia que en la medida que los países sur-orientales de europa se acercan geográficamente al occidente o el norte de europa sus frecuencias de resistencia antibiótica es menor. Esto debe ser comprobado a través de más estudios geopoblacionales. Sin embargo, permite lanzar supuestos respecto a la conciencia acerca del uso racional de los antibióticos

en occidente y también a las mayores prácticas de regulación y control que se han adoptado en muchos países de europa occidental para contener esta problemática de caracter mundial.