Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

Validación

Validación

“ The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes, 1988.

Validaciones

Relativa: evalúa la estructura de clusters variando los distintos parámetros del algoritmo que se haya utilizado. En general se utiliza para determinar el número “óptimo”de clusters.

Externa: compara los resultados que hayamos obtenido del análisis de cluster con información externa. Si conocemos de antemano las etiquetas correctas, este tipo de enfoque se utiliza para comparar entre distintos algoritmos. En otros casos la información externa viene dada por un experto.

Validaciones

Interna: usa la información que genera el proceso de clusterización para evaluar la bondad de la estructura que hayamos conseguido sin utilizar ninguna información externa.

Estabilidad: evalúa la consistencia de los clusters comparando con ligeras modificaciones del dataset.

Enfoque

- Uso de información externa.
- Exploración visual
- Estabilidad
- Índices internos para la validación
- Sensibilidad y comparación de diferentes particiones en el mismo dataset. ¿Cuántos Clusters?

Información externa

Podemos tener distintos tipos de información externa:

- Variables que sabemos que deben estar relacionadas a la partición:
  - Tener las etiquetas de antemano.
  - Conocer una clasificación relativa. Por ejemplo, (países ricos y pobres).
- Tener una variable de antemano que queremos que sea predecida o explicada por la partición.
- Asesoramiento de un experto en el tema con el que estemos trabajando.

Información externa

En el caso de conocer las etiquetas de antemano, el problema de clustering está resuelto.

Sin embargo, este tipo de datasets se utilizan para construir benchmarks contra los cuales podemos comparar diferentes métodos. Los índices más utilizados para esta tarea son:

- Tasa de clasificación correcta (CCR).
- Rand index.
- Adjusted Rand Index.

Exploración visual

Tratamos de detectar la calidad del análisis a partir de una interpretación gráfica de los clusters.

Afortunadamente el ojo humano tiene una gran capacidad para detectar patrones y entender las características de los métodos.

Exploración visual

Tratamos de detectar la calidad del análisis a partir de una interpretación gráfica de los clusters.

Afortunadamente el ojo humano tiene una gran capacidad para detectar patrones y entender las características de los métodos.

El gran problema es que no se puede visualizar cuando la dimensión de los datos es mayor a 3. Tenemos algunas herramientas para tener representaciones en dimensiones menores:

- Proyecciones (PCA y derivados).
- Multidimensional Scaling.
- Heatmaps.

Exploración visual

Iris dataset (ejemplo de juguete), contiene distinas medidas para tres especies distintas de plantas iris.

- sepal length (cm).
- sepal width (cm).
- petal length (cm).
- petal width (cm).
- clase:
  - Iris Setosa.
  - Iris Versicolour.
  - Iris Virginica.

Exploración visual

Figura 1: Iris Versicolor

Exploración visual

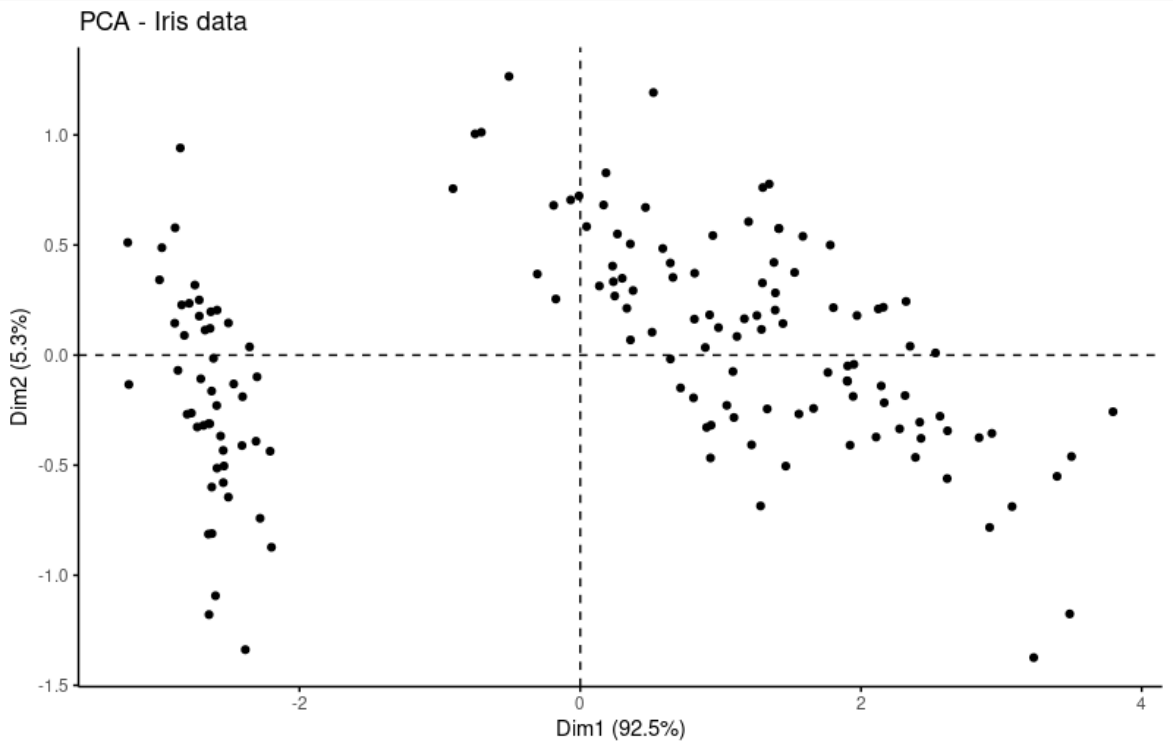


Figura 2: Primeras dos componentes principales

Exploración visual

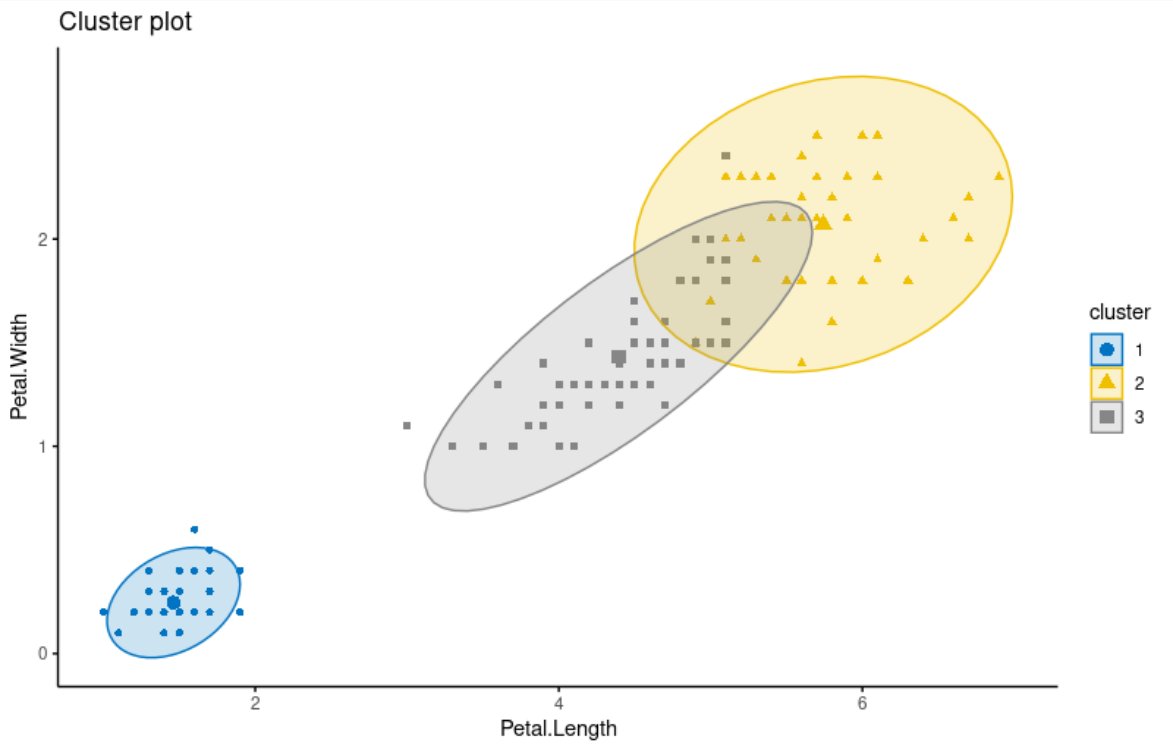


Figura 3: kmeans 3 clusters

Exploración visual

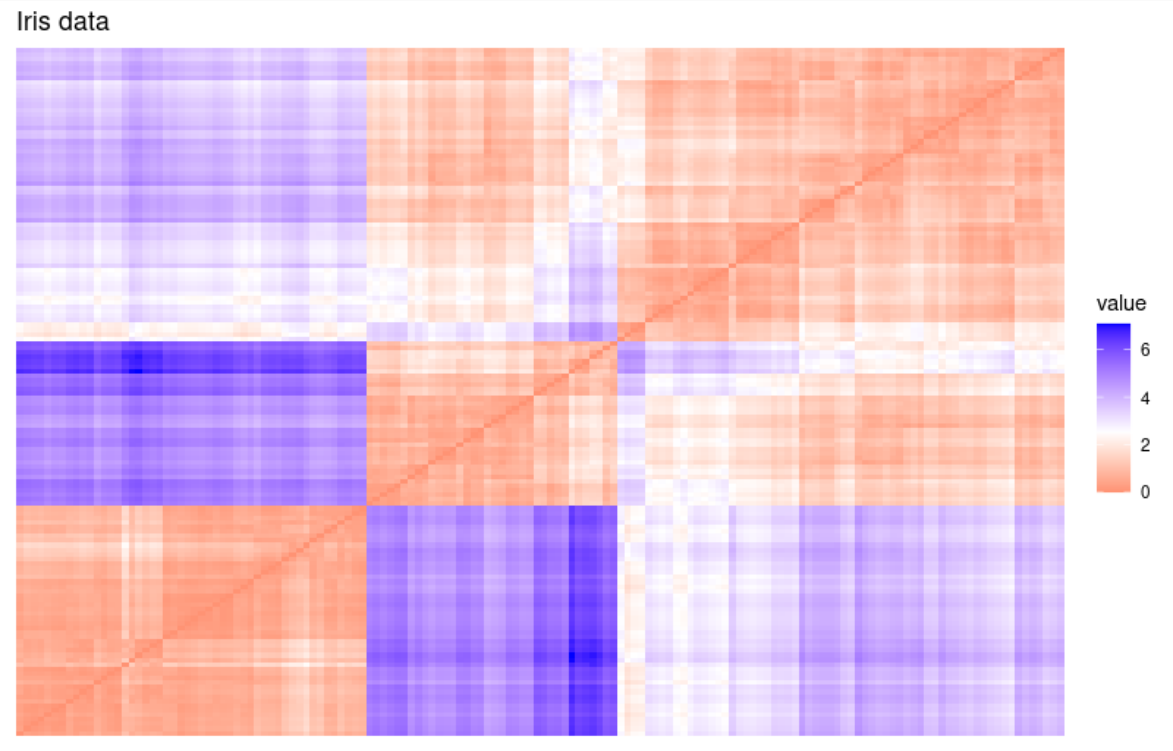


Figura 4: Heatmap basado en la disimilaridad

Estabilidad

En general una evaluación de estabilidad se basa:

- Generar de alguna manera nuevos datasets a partir del original.
- Clusterizar estos nuevos datasets.
- Definir un estadístico que mida que tan parecidos son los nuevos clusters con respecto a los orginales.
- Concluir que si son similares, son estables!

Observación: tener estabilidad es una buena propiedad, pero clusters estables no garantizan “buenos” clusters.

Algoritmo Hening

Usamos el coeficiente de Jaccard que es un forma simple de medir similaridad entre dos conjuntos (no requiere de una métrica).

$$\gamma(C, D) = \frac{|C \cap D|}{|C \cup D|}$$

- Generar B muestras bootstrap del dataset. También es posible utilizar otro tipo de resampleo aunque cambia un poco la interpretación.
- Aplicar a cada muestra bootstrap el método de cluster que hayamos elegido.
- Para cada $C \in \mathcal{C}$ registrar $m_{b,C} = \text{máx}_{D \neq C}\, \gamma(C, D)$.
- Calcular $\overline{\gamma} = \frac{1}{|B|} \sum_{b \in B} m_{b,C}$.

Algoritmo Hening

Podemos tomar el 0,5 como corte de estabilidad,

- Considerar los clusters con $\gamma < 0{,}5$ como disueltos.
- Tomar como estables los que tengan $\overline{\gamma} >> 0{,}5$.

Consideremos este ejemplo,

$\overline{\gamma_1}$	$\overline{\gamma_2}$	$\overline{\gamma_3}$	$\overline{\gamma_4}$
0.568	0.773	0.500	0.487
$\overline{\gamma_5}$	$\overline{\gamma_6}$	$\overline{\gamma_7}$	$\overline{\gamma_8}$
0.962	0.927	0.908	0.780
$\overline{\gamma_9}$			
0.885			

Algoritmo Hening

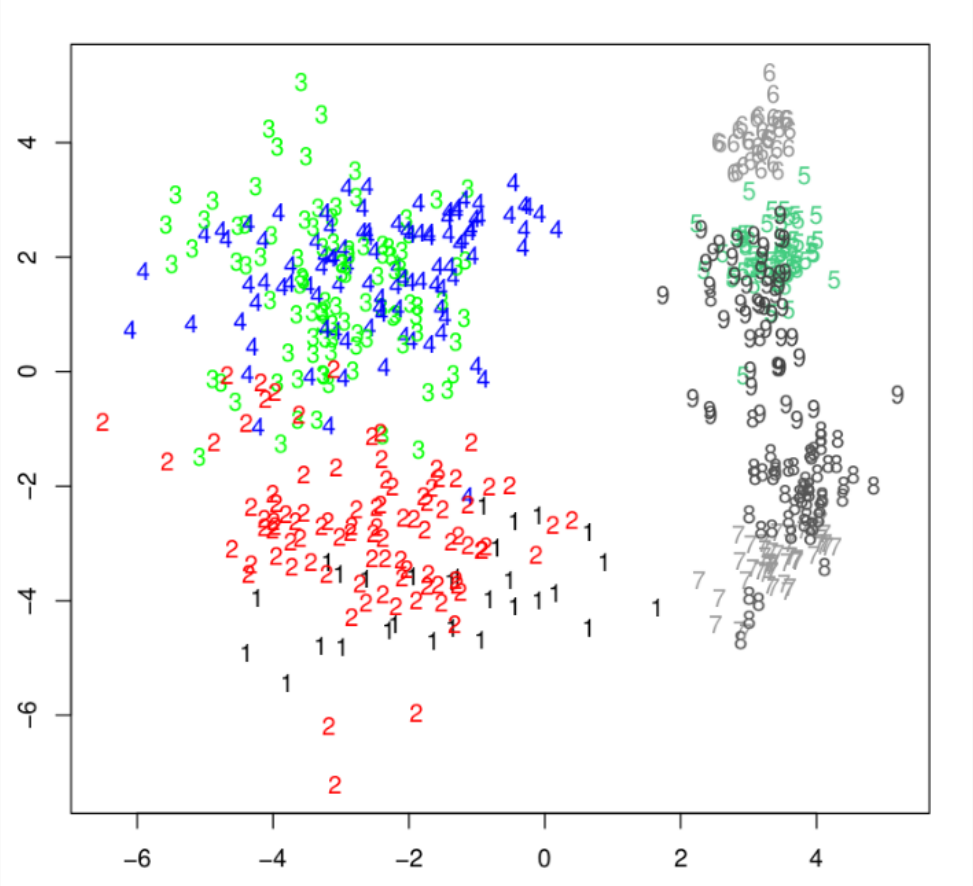


Figura 5: Ejemplo con 9 grupos

Validación Interna

Características

Cohesión

Mide que tan cercanos son los objetos que se encuentran dentro del mismo cluster. Por ejemplo, tener baja varianza intra-grupos es un buen indicador de la cohesión si nuestro algoritmo lo basamos en distancias.

Características

Cohesión

Mide que tan cercanos son los objetos que se encuentran dentro del mismo cluster. Por ejemplo, tener baja varianza intra-grupos es un buen indicador de la cohesión si nuestro algoritmo lo basamos en distancias.

"Lo que se parece debe estar cerca".

Características

Separación

Mide que tan separados están los clusters entre sí. En general estos índices utilizan la distancia entre los centroides de los clusters o la distancia entre objetos que se encuentran en distintos clusters.

Características

Separación

Mide que tan separados están los clusters entre sí. En general estos índices utilizan la distancia entre los centroides de los clusters o la distancia entre objetos que se encuentran en distintos clusters.

"Lo que no se parece debe estar lejos".

Características

Conexión

Mide la relación entre un punto y sus vecinos más cercanos, es decir, trata de darnos una noción de un entorno alrededor de cada objecto de cada cluster.

Características

Conexión

Mide la relación entre un punto y sus vecinos más cercanos, es decir, trata de darnos una noción de un entorno alrededor de cada objecto de cada cluster.

"Dime con quién andas y te diré quién eres".

Silohuette

El índice de Silohuette compara una medida de cohesión con una medida de separación.

En general muchos de los índices de validación interna se construyen a partir de estas medidas.

Utilicemos el Silohuette como ejemplo para entender estos índices.

Silohuette

¿Cómo lo calculamos?

Supongamos que tenemos nuestros datos $X = \{x_1, \ldots, x_n\}$ y consideramos una partición $\mathcal{C} = \{C_1, \ldots, C_K\}$.

Cohesión: para cada x_i calculo la disimilaridad media con respecto a los datos que están en el mismo cluster. Es decir, si x_i está en el cluster C, defino:

$$a(i) = \frac{1}{|C| - 1} \sum_{x_j \in C, x_j \neq x_i} d(x_i, x_j).$$

Silohuette

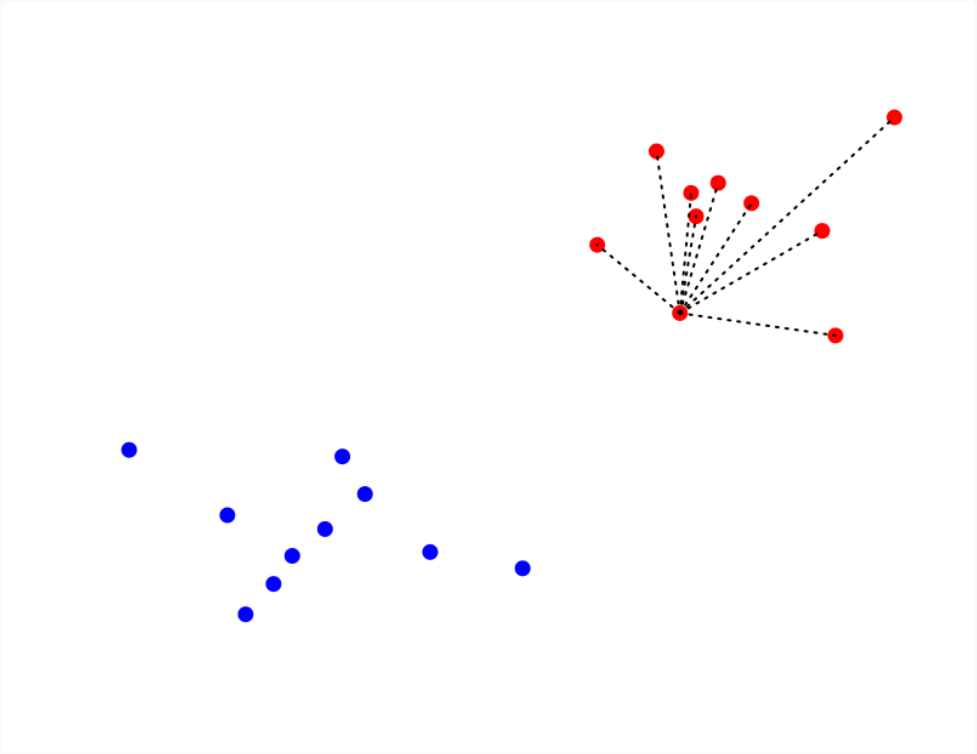


Figura 6: Cohesión

Silohuette

Separación: tengo que medir la separación con respecto a todos los otros clusters donde no está x_i. Es decir,

$$b(i) = \min_{B \in \mathcal{C}:\ B \cap C = \emptyset} \frac{1}{|B|} \sum_{z \in B} d(x_i, z).$$

Silohuette

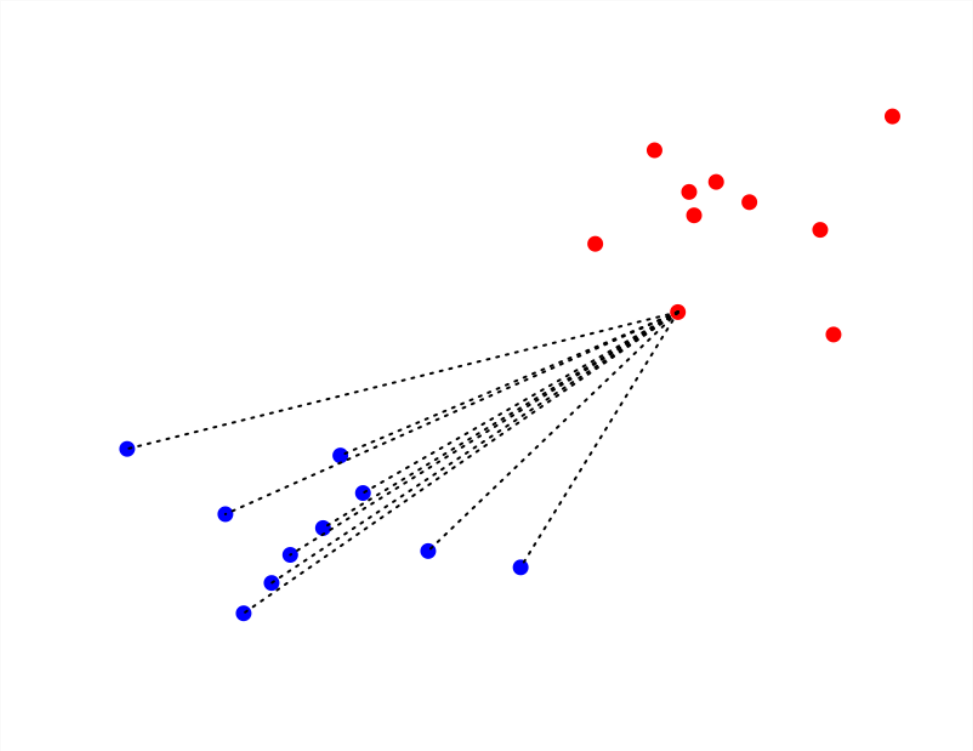


Figura 7: SEPARACION

Silohuette

Ahora, para cada dato x_i, $1 \leq i \leq n$ tengo una medida de cohesión y separación.

Se define el coeficiente de Silohuete para x_i como

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \ 1 \leq i \leq n.$$

Observación: $-1 \leq s(i) \leq 1, \ \forall \ i.$

Silohuette

Ahora, para cada dato x_i, $1 \leq i \leq n$ tengo una medida de cohesión y separación.

Se define el coeficiente de Silohuete para x_i como

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \ 1 \leq i \leq n.$$

Observación: $-1 \leq s(i) \leq 1, \ \forall \ i.$

Además para cluster $C \in \mathcal{C}$ podemos definir su coeficiente de Silhouette como el promedio de los coeficientes de cada punto. También podemos obtener el Silhouette medio para cada partición promediando todos los coeficientes.

Silohuette

Interpretación: notar que es un índice muy intuitivo,

- Un valor cercano a -1 no dicen que $b(i)$ es mucho más chico que $a(i)$. Con lo cual ese dato está alejado de sus compañeros de cluster y no tan alejado de los elementos que están en los clusters a los cual no pertenece.
- Un valor cercano a 1, contrariamente nos dice lo opuesto. Es decir, que tiene buena separación y cohesión.

Silohuette

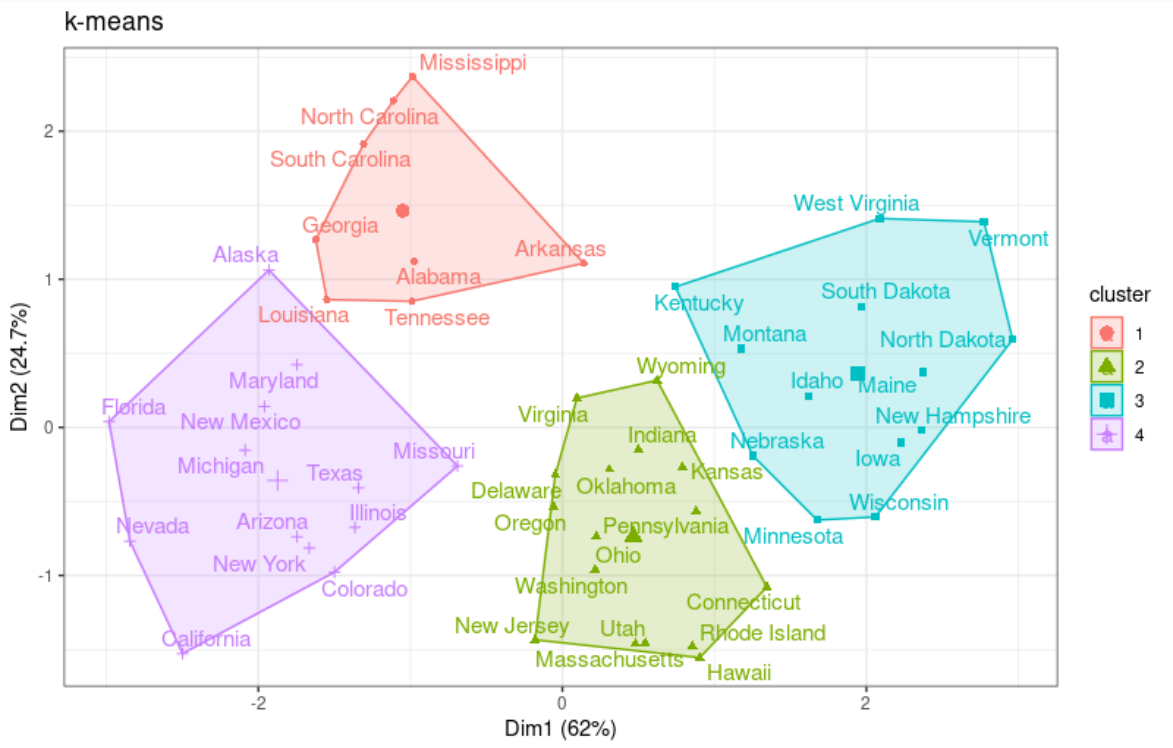


Figura 8: Dataset Arrestos EEUU

Silohuette

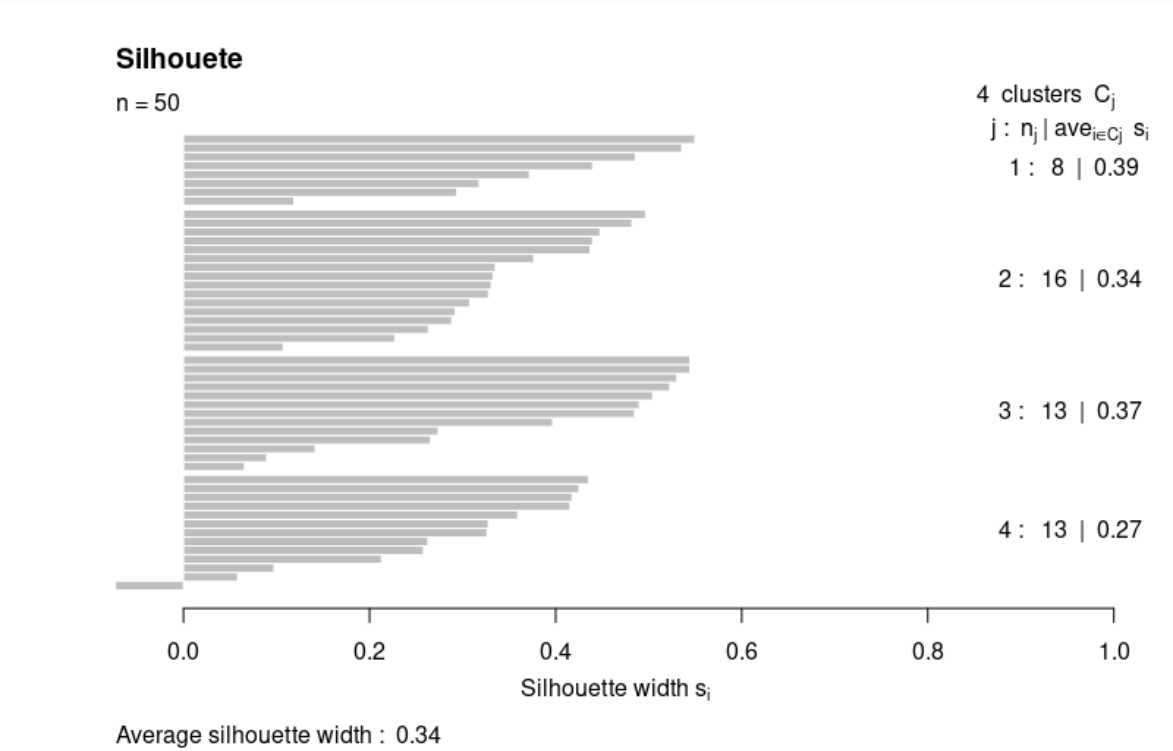


Figura 9: Silhouetes

Dunn-Index

Es una medida para comparar la separación y la cohesión de los grupos. También es útil como veremos unas diapositivas más adelante para responder a la pregunta ¿cuántos clusters?

El valor depende de cómo se mide, pero una fórmula general es

$$\frac{\text{mín}_{1\leq i<j\leq K}\, Dist(C_j, C_j)}{\text{máx}_{1\leq i\leq K}\, \text{diámetro}(C_i)}.$$

Ci,Cj

Antes de usar el Dunn-Index chequear cómo define la implementación del algoritmo cada cantidad.

¿Cuántos Clusters?

Es la pregunta del millón!!!

¿Cuántos Clusters?

Es la pregunta del millón!!!

No existe una forma cerrada para la elección del número optimo de grupos.

¿Cuántos Clusters?

Es la pregunta del millón!!!

No existe una forma cerrada para la elección del número optimo de grupos.

Muchas veces el número “óptimo”de clusters depende del método o algoritmo utilizado.

¿Cuántos Clusters?

Es la pregunta del millón!!!

No existe una forma cerrada para la elección del número optimo de grupos.

Muchas veces el número “óptimo”de clusters depende del método o algoritmo utilizado.

Es muy válido utilizar información externa o el criterio de un experto en este caso.

Vamos a ensuciarnos las manos

Algunos métodos

- Elbow
- Gap Statistic
- Max Mean Silhouette
- Maximizar Dunn-Index

Elbow

Es un método heurístico que se basa en graficar la varianza intra-clusters contra el número de clusters.

Elbow

Es un método heurístico que se basa en graficar la varianza intra-clusters contra el número de clusters.

La varianza intra-clusters es decreciente con respecto al número de grupos que se toman en la partición.

El método Elbow se basa en encontrar justamente el “codo de la curva”, es decir, el valor del eje horizontal para el cual el decrecimiento comienza a ser más lento.

Elbow

Es un método heurístico que se basa en graficar la varianza intra-clusters contra el número de clusters.

La varianza intra-clusters es decreciente con respecto al número de grupos que se toman en la partición.

El método Elbow se basa en encontrar justamente el “codo de la curva”, es decir, el valor del eje horizontal para el cual el decrecimiento comienza a ser más lento.

Veamos un ejemplo...

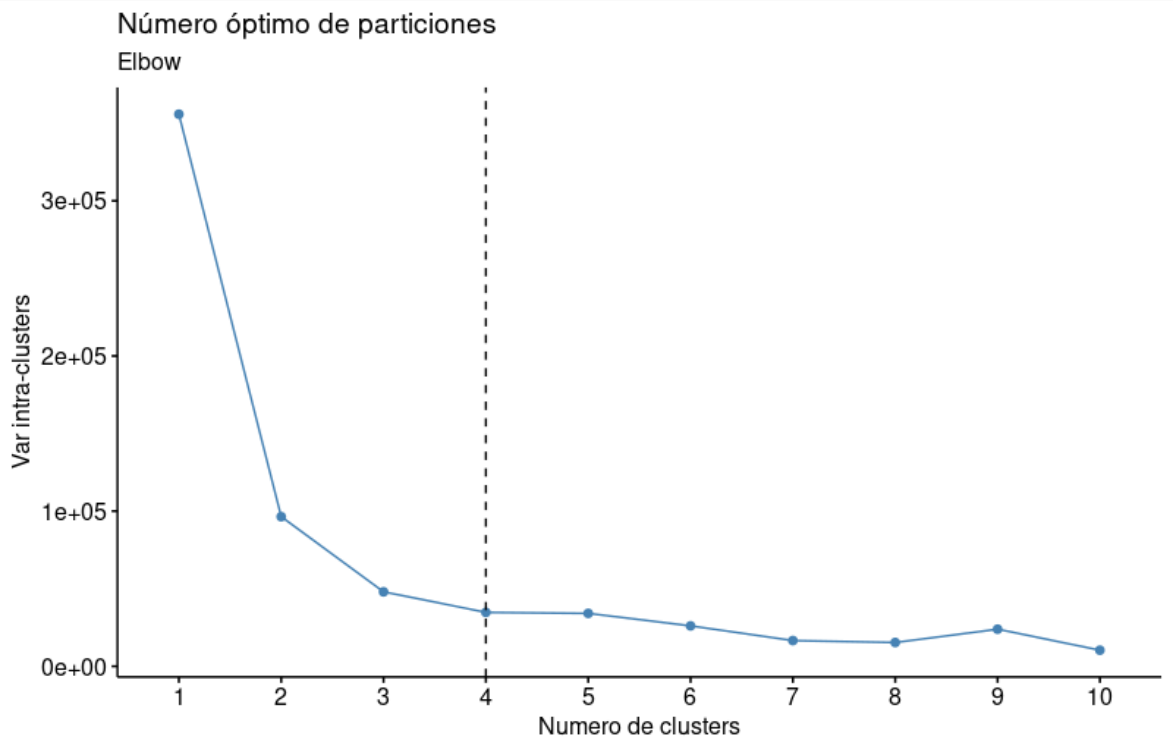


Figura 10: Elbow: 4 clusters

Silhouette

Para cada partición $\mathcal{C}$ de K clusters, calcular su Silhouette medio, S_K.

Luego, tomar como número óptimo de clusters la partición que tenga máximo Silhouette medio.

$$K_{opt} = \arg\max_{K \geq 2} S_K.$$

Silhouette

Para cada partición $\mathcal{C}$ de K clusters, calcular su Silhouette medio, S_K.

Luego, tomar como número óptimo de clusters la partición que tenga máximo Silhouette medio.

$$K_{opt} = \arg\max_{K \geq 2} S_K.$$

Observación: con la misma lógica podemos utilizar el Dunn-Index.

USArrest

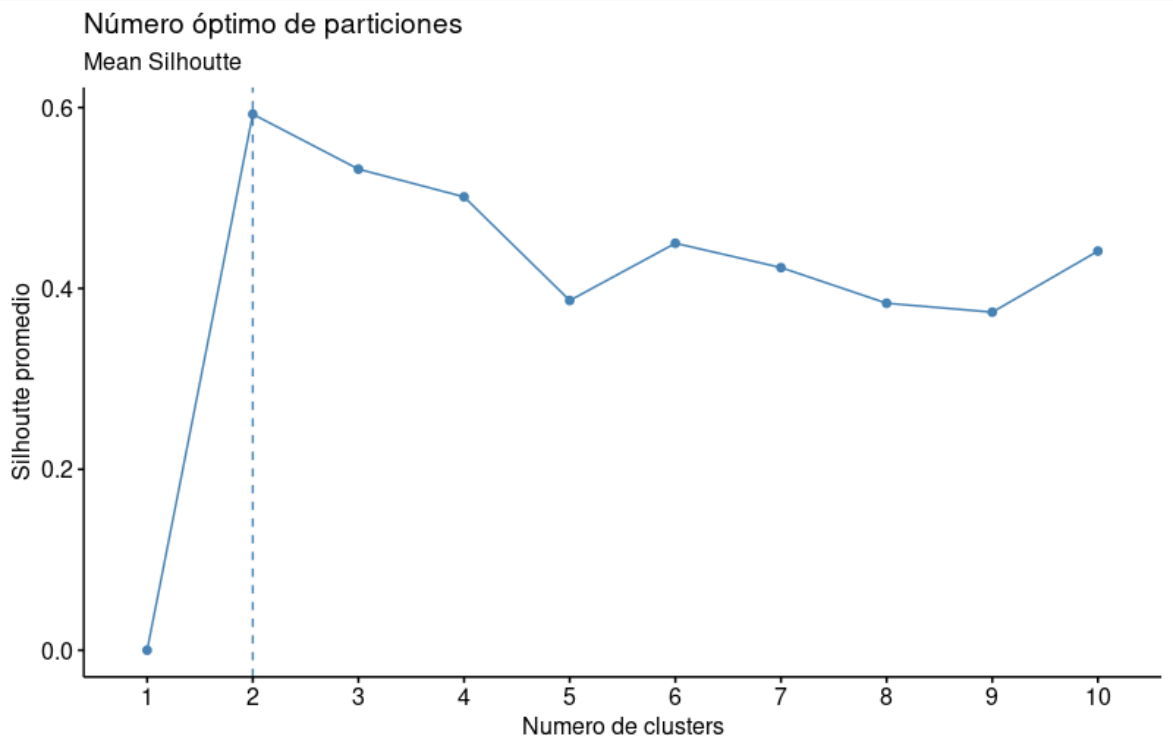


Figura 11: Max Mean Silhouette: 2 clusters

Gap

Idea: comparar la distribución del conjunto de datos con el que estoy trabajando con una distribución donde yo sepa de antemano que no hay ningún grupo.

Para cada partición $\mathcal{C}$ de tamaño K de D, llamemos W_K a su varianza intra-clusters.

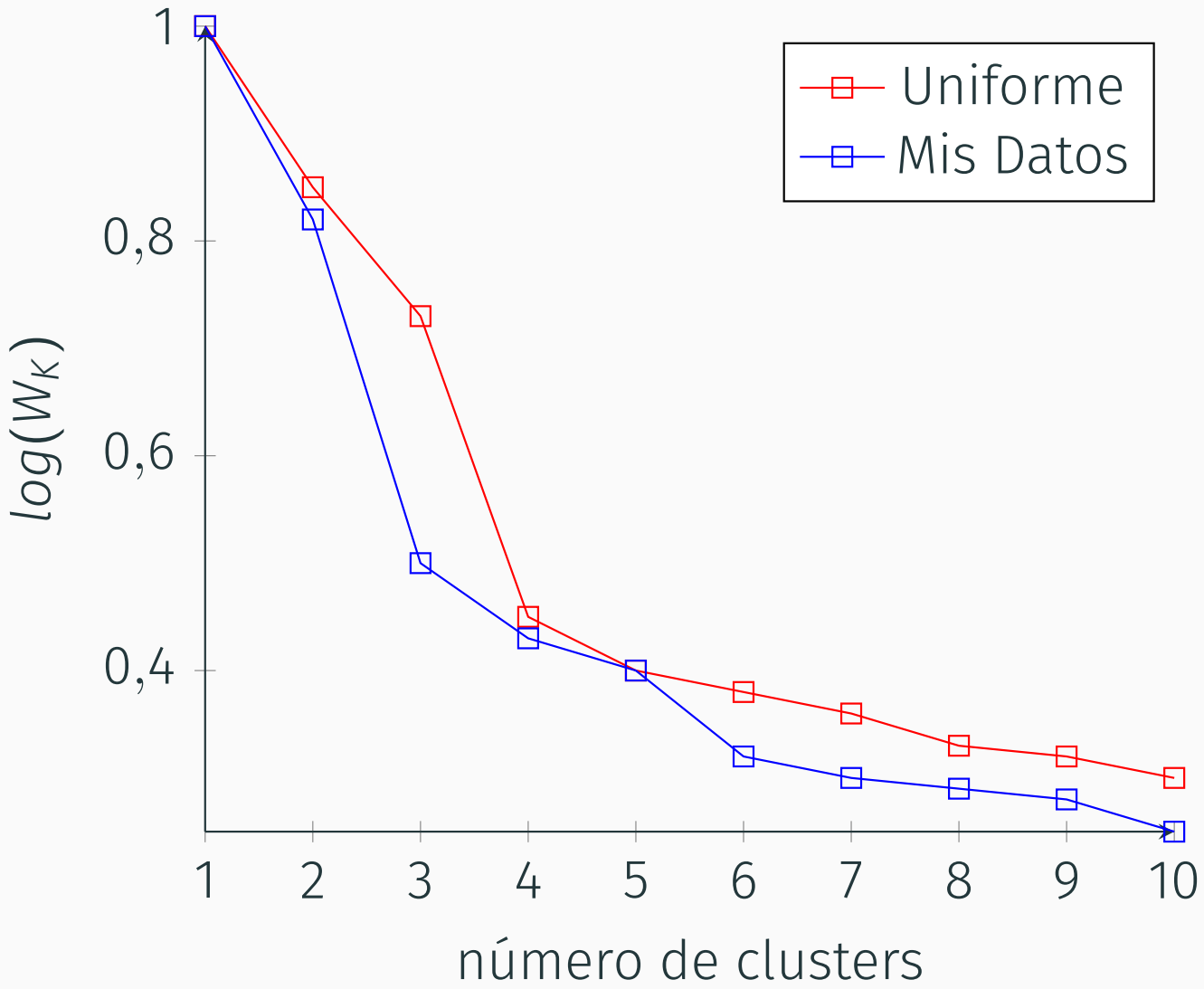
Gap

El método compara la curva de $log(W_K)$ con la misma curva que se obtendría con datos uniformemente distribuídos sobre un rectángulo que contiene a mis datos.

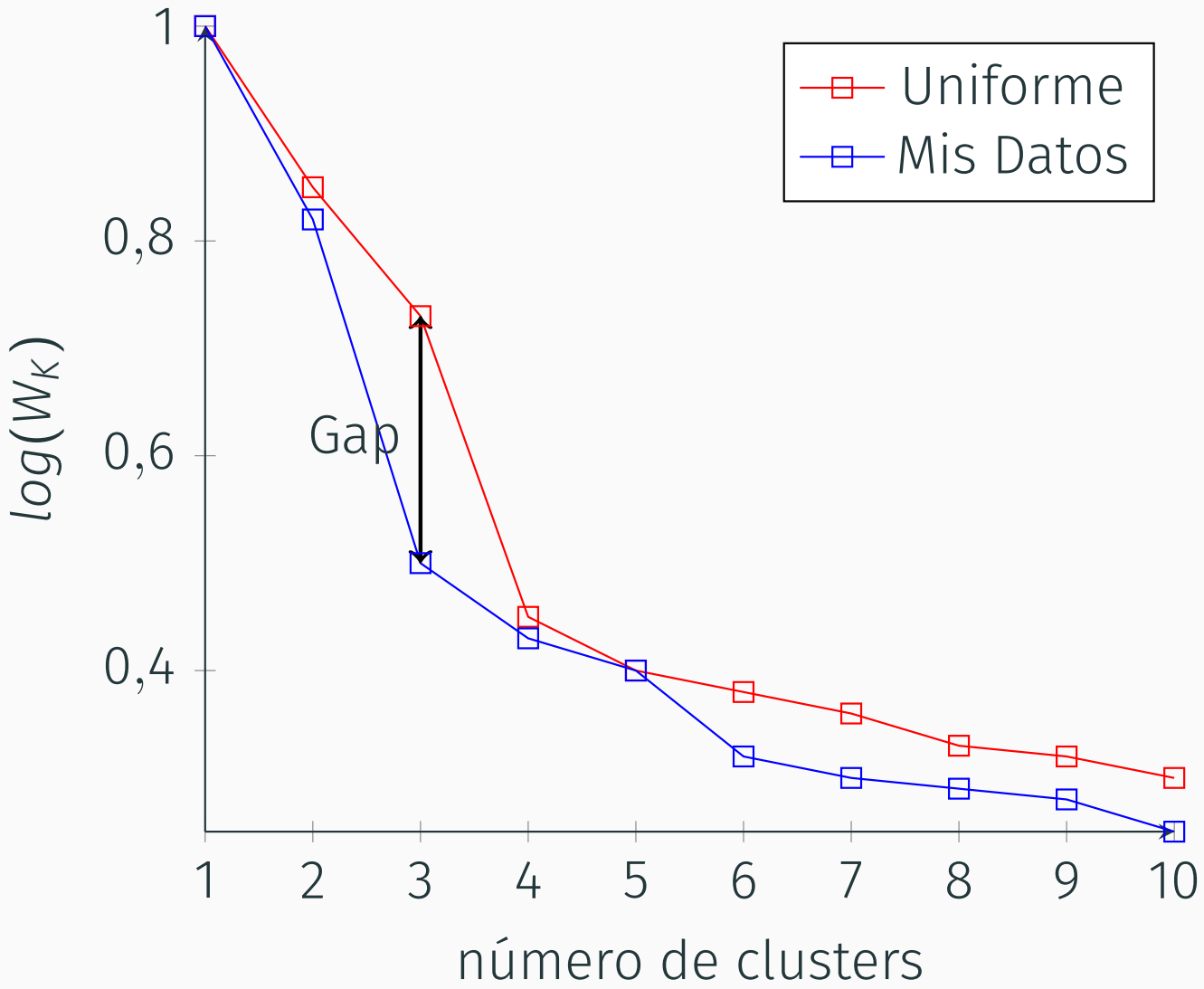
Estima el número óptimo de grupos como el K donde ambas curvas están más separadas. Por eso mismo recibe el nombre de “Gap”.

Observación: a diferencia de otros métodos tiene permitido estimar el número óptimo de clusters como 1.

Gap


1
0,8
0,6
0,4
$log(W_K)$
Uniforme
Mis Datos
1
2
3
4
5
6
7
8
9
10
número de clusters

Gap


1
0,8
0,6
0,4
$log(W_K)$
Gap
Uniforme
Mis Datos
1
2
3
4
5
6
7
8
9
10
número de clusters

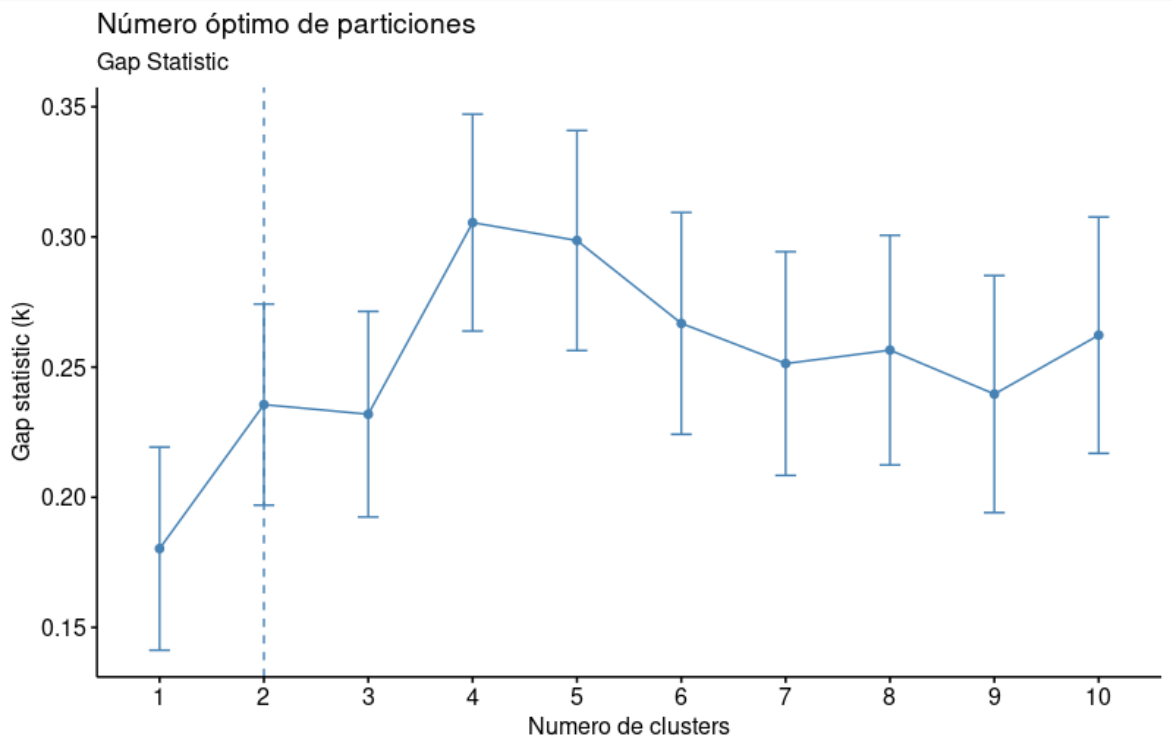


Figura 12: Gap Statistic: 2 clusters

USArrest 4 clusters

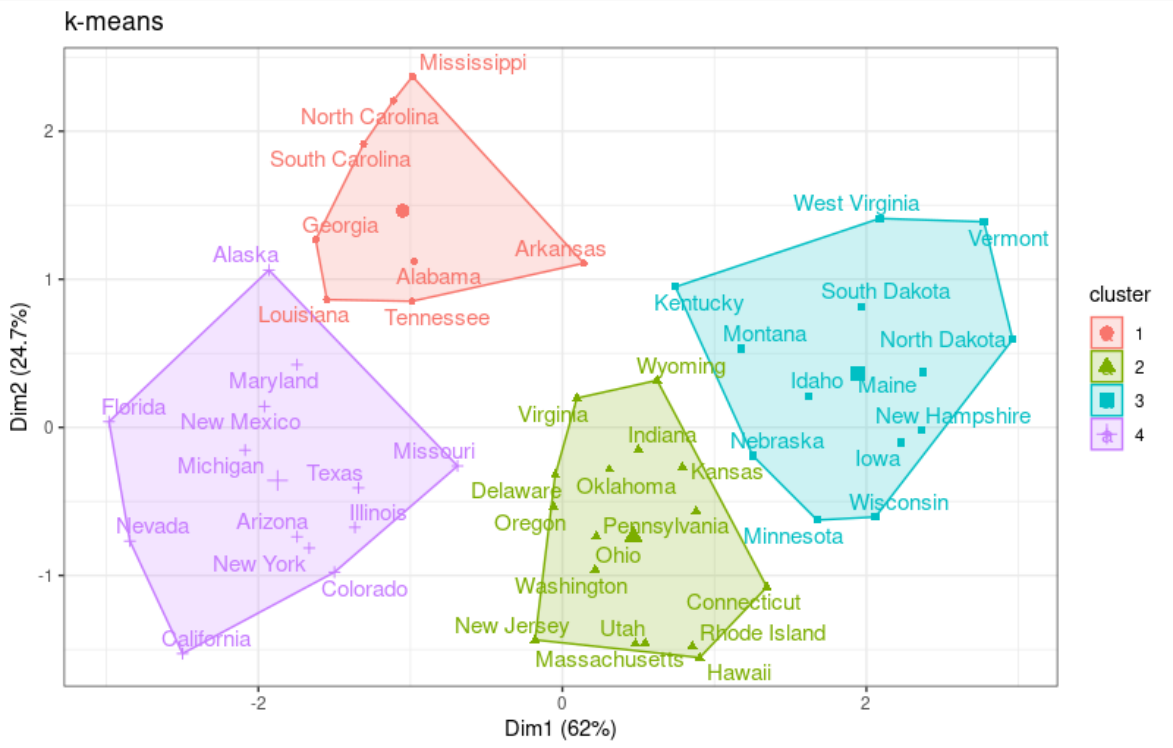


Figura 13: kmeans

USArrest 2 clusters

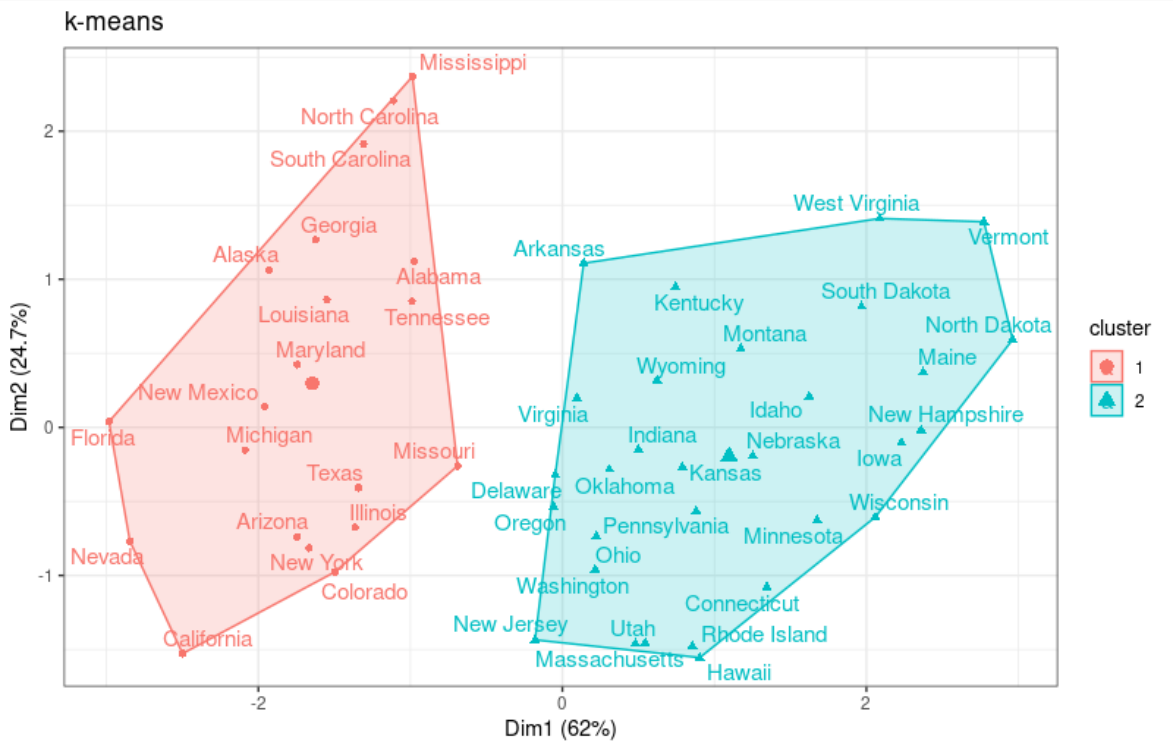


Figura 14: kmeans

USArrest ¿Clusters?

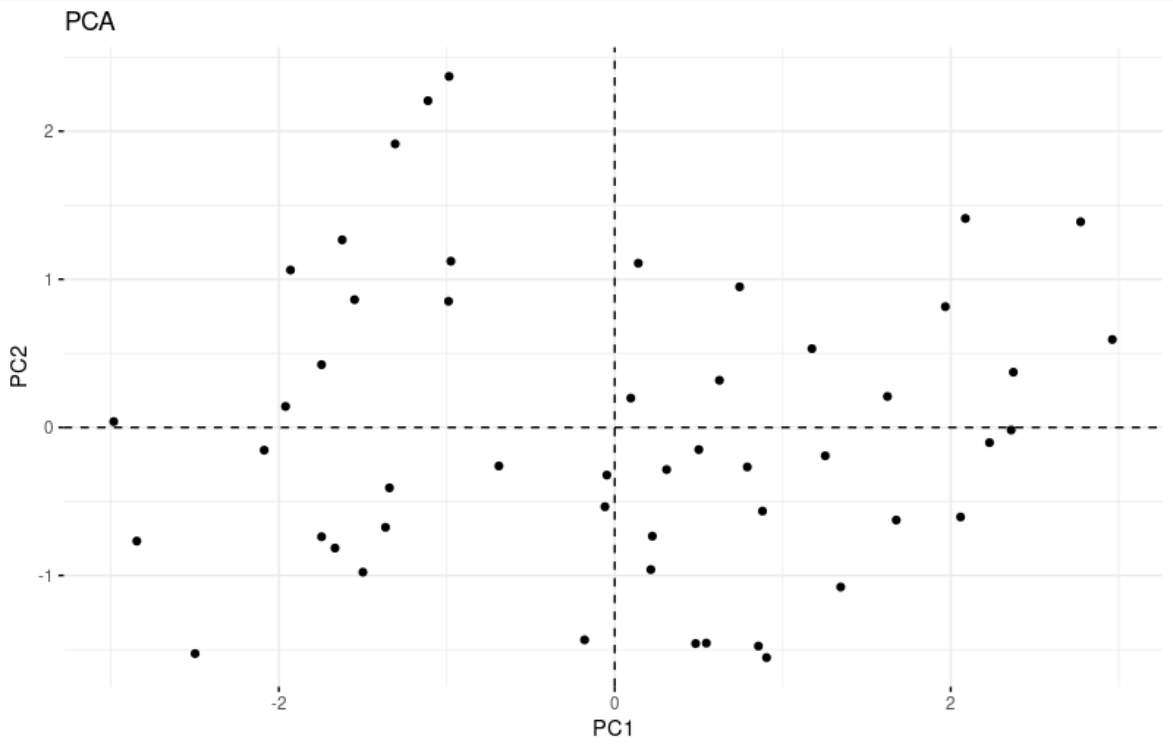


Figura 15: ¿Estamos viendo estructura o ruido en USArrest?

COFFEE BREAK!