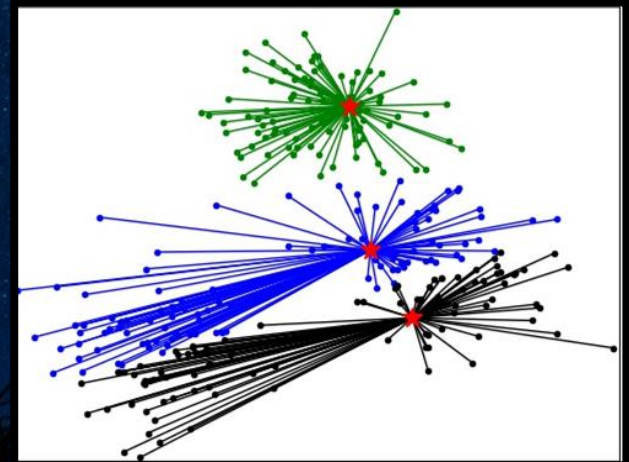




K-means

# K-means

- El algoritmo K-means (K-medias) es un algoritmo partitivo.
- Uno de los más utilizados y más simples de implementar.
- Dado un conjunto de datos y K centroides: encontrar la posición de los centros que minimicen la suma de distancias.
- Cada centroide representa un *cluster*.
- Minimiza la suma de distancias cuadradas.






# K-means – Algoritmo de Lloyd

## Inicialización

Crear aleatoriamente  $k$  centros (dentro del dominio)

## Iterar hasta converger (no cambian las asignaciones)

1. Para cada patrón en el conjunto de datos:  
calcular la distancia a cada centroide  
asignar el patrón al *cluster* más cercano
2. Para cada *cluster*:  
mover el centroide al nuevo punto medio



Etapas de asignación  
de cluster



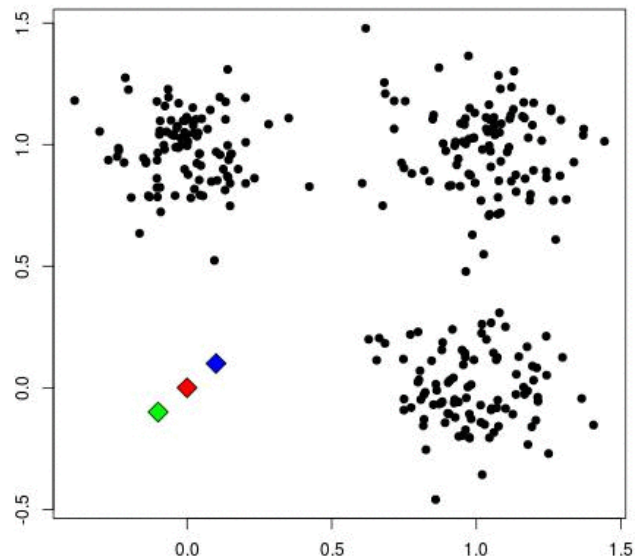
Etapas de actualización  
del centroide

## Para usar el algoritmo con un nuevo patrón:

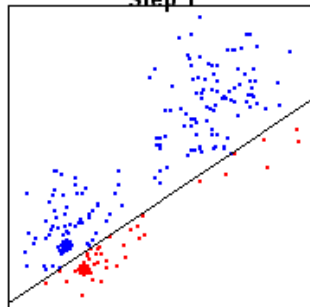
1. Calcular la distancia del patrón a cada centroide.
2. Asignar el patrón al *cluster* con menor distancia

# K-means

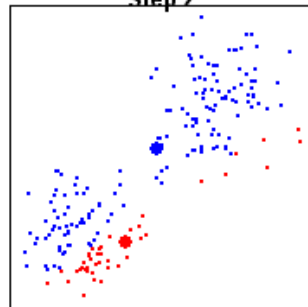
Start!



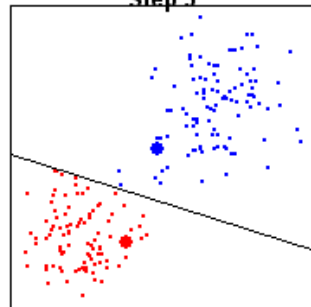
Step 1



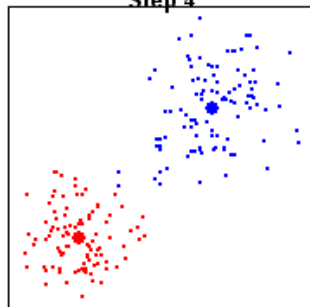
Step 2



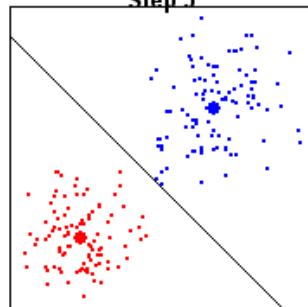
Step 3



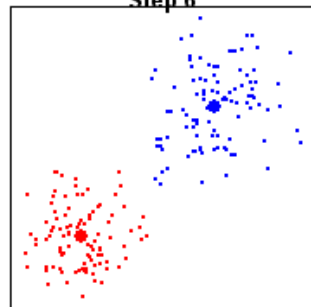
Step 4



Step 5



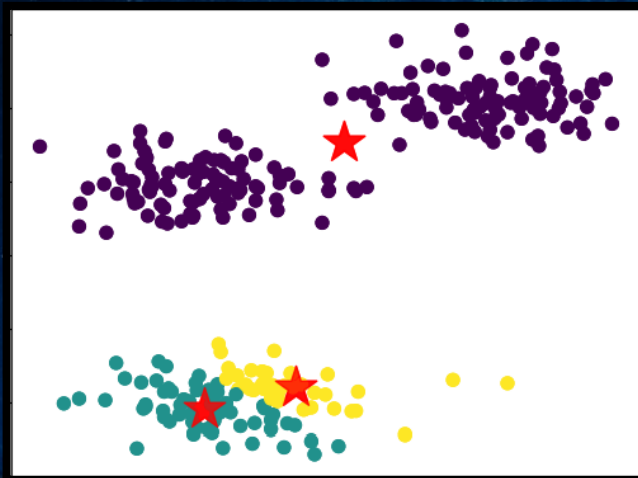
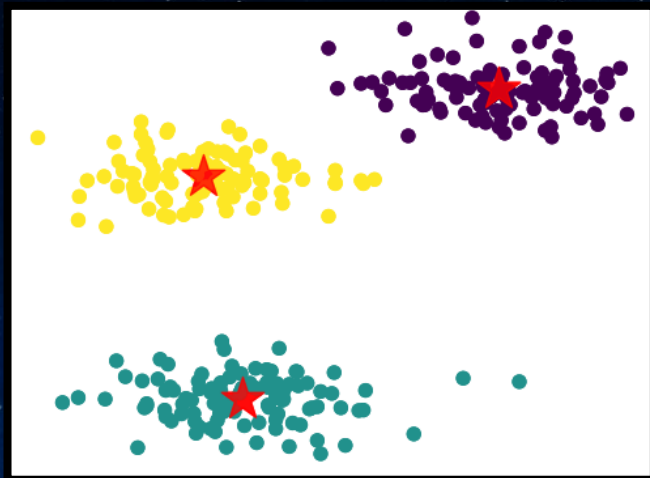
Step 6



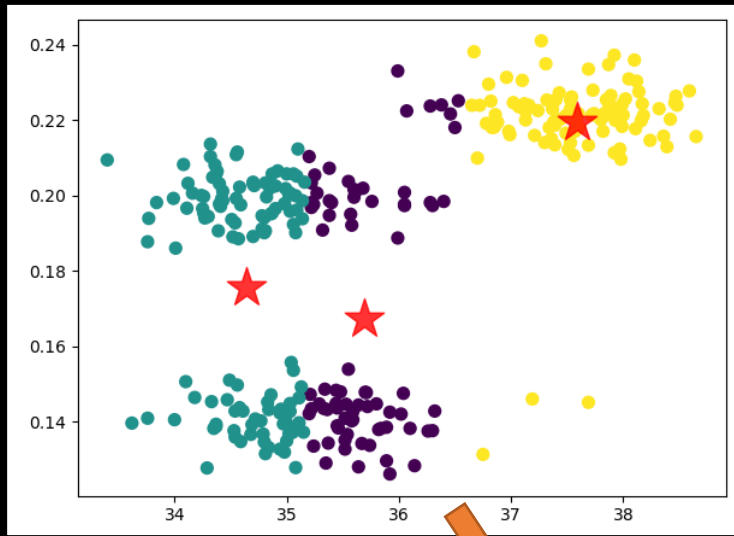
# K-means – Mínimos locales

Cada paso del algoritmo de Lloyd decrementa la SSD

- No es una función convexa si  $K > 2$
- Mínimos locales
- Soluciones: kmeans++, varias iteraciones, etc.

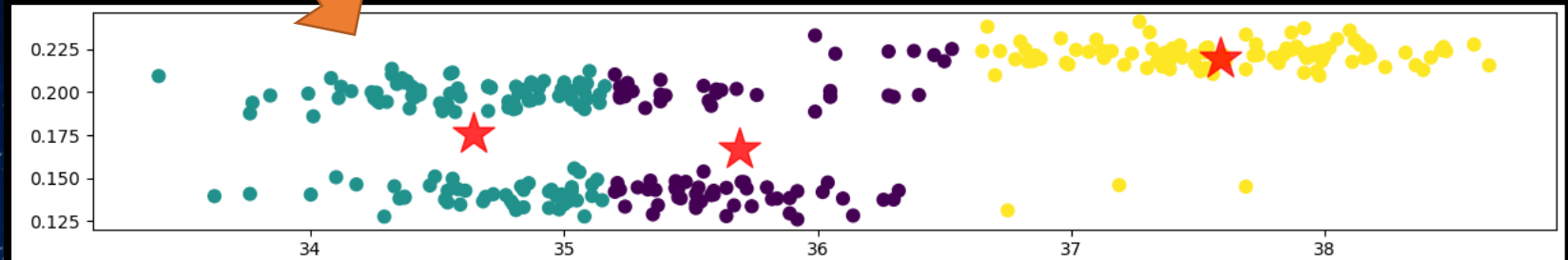


# K-means – Normalización



K-means es sensible a la escala de los datos

- Utiliza medida de distancia
- Es importante normalizar





# K-means en SciKit-Learn

```
from sklearn.cluster import Kmeans
```

```
Df = pd.read_csv('calabazas.csv')
```

```
K= 5
```

```
kmeans_model = KMeans(n_clusters=k)
```

```
kmeans.fit(Df)
```

```
centers = kmeans.cluster_centers_
```

```
kmeans_labels = kmeans.predict(Df)
```

Crea el modelo  
con K clusters

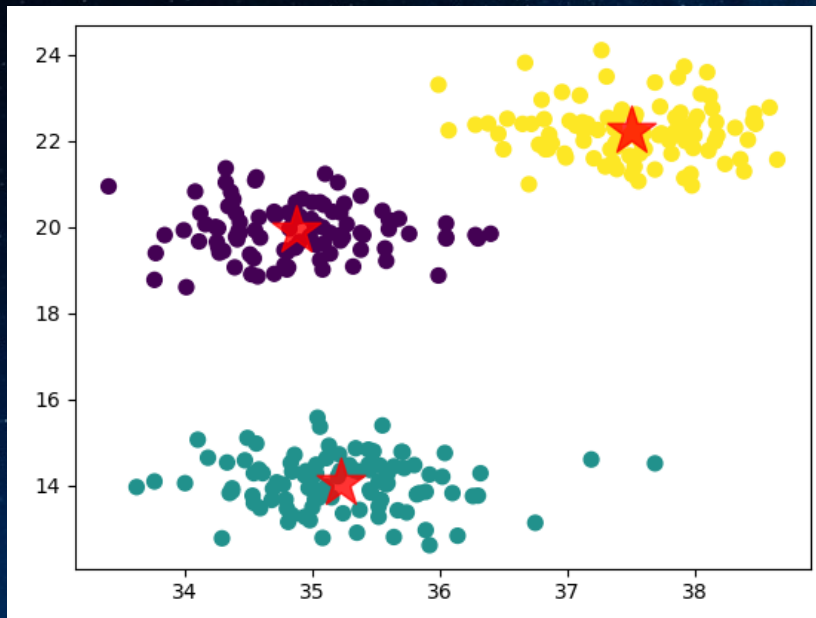
Entrena el  
modelo

Centroides

Etiquetar  
nuevos datos

# K-means en SciKit-Learn

```
plt.scatter(data[:, 0], data[:, 1], c=kmeans_labels)  
# visualizar centros  
plt.scatter(centers[:, 0], centers[:, 1], marker='*', c='red')
```





# Ejemplo. WHO\_life\_expectancy

```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler()
```

```
data= scaler.fit_transform(data)
```

**Normalizamos con Min-Max (más interpretable)**

```
k= 2
```

```
kmeans = KMeans(n_clusters=k).fit(data)
```

```
centers = kmeans.cluster_centers_
```

**Creamos 2 grupos**

**Desnormalizamos los centros y los guardamos**

```
df2 = pd.DataFrame(scaler.inverse_transform(centers), columns= columnas)
```

```
df2.to_excel("centros_paises2.xlsx")
```

# Ejemplo. WHO\_life\_expectancy

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
data= scaler.fit_transform(data)

k= 2
kmeans = KMeans(n_clusters=k).fit(data)
centers = kmeans.cluster_centers_

df2 = pd.DataFrame(scaler.inverse_transform(centers),
columns= columnas)
df2.to_excel("centros_paises2.xlsx")
```

## Centro 1:

- Países con baja esperanza de vida.
- Poco consumo de alcohol.
- Gran población.

VARIABLE	Centro 1	Centro 2
Life expectancy	61,92	74,56
Adult Mortality	239,29	117,59
infant deaths	67,73	7,49
Alcohol	2,31	6,12
percentage expenditure	129,20	1.104,86
Hepatitis B	67,91	87,27
Measles	3.770,07	1.123,49
BMI	23,10	48,83
under-five deaths	94,01	8,75
Polio	71,66	92,04
Total expenditure	5,41	6,35
Diphtheria	72,35	92,56
HIV/AIDS	4,45	0,23
GDP	1.323,41	8.588,30
Population	22.020.898,75	9.405.495,89
thinness 1-19 years	7,89	2,69
thinness 5-9 years	8,05	2,67
Income composition of resources	0,48	0,74
Schooling	9,86	13,73
SOPORTE	685,00	964,00

# Ejemplo. WHO\_life\_expectancy

## 10 Clusters

VARIABLE	Centro 1	Centro 2	Centro 3	Centro 4	Centro 5	Centro 6	Centro 7	Centro 8	Centro 9	Centro 10
Life expectancy	61,87	77,80	59,81	72,46	75,48	66,00	68,77	66,65	72,45	51,57
Adult Mortality	211,23	98,89	248,47	128,42	111,60	116,64	164,39	197,38	136,32	442,79
infant deaths	45,21	2,67	78,35	7,94	18,62	1.268,82	24,49	44,88	24,02	38,83
Alcohol	1,07	10,27	2,86	3,57	6,88	2,26	1,76	3,07	4,46	4,75
percentage expenditure	70,84	2.438,23	163,04	311,83	1.507,03	41,48	185,37	267,86	418,87	173,79
Hepatitis B	79,40	91,17	28,13	92,56	15,28	30,27	66,63	77,39	93,96	79,23
Measles	1.636,76	475,86	3.920,45	494,18	2.675,83	43.188,55	3.359,44	1.662,96	4.860,15	2.946,43
BMI	18,44	57,53	23,40	52,12	44,96	15,45	44,44	36,33	18,70	23,31
under-five deaths	63,03	3,10	120,58	9,27	22,99	1.681,82	30,40	61,84	28,70	59,68
Polio	82,43	95,05	38,31	93,23	85,48	72,55	75,21	8,38	95,39	82,98
Total expenditure	5,23	7,27	5,11	6,11	5,78	4,36	4,97	5,75	5,45	6,76
Diphtheria	82,58	94,80	22,94	93,31	82,07	68,09	65,39	76,75	95,35	83,69
HIV/AIDS	1,72	0,11	4,33	0,25	0,33	0,25	0,52	2,18	0,49	18,78
GDP	733,24	17.706,72	1.644,12	3.230,19	10.439,17	900,01	1.949,52	2.225,40	4.058,43	1.674,57
Population	11.834.394,35	12.345.785,23	11.722.462,07	8.579.982,59	12.858.878,62	594.387.175,45	15.770.977,61	14.944.287,71	8.080.730,07	7.401.515,42
thinness 1-19 years	9,65	1,45	7,42	2,80	2,35	27,00	2,72	4,11	6,19	7,44
thinness 5-9 years	9,42	1,46	7,23	2,77	2,46	27,82	3,08	4,73	6,48	7,70
Income composition of resources	0,42	0,83	0,47	0,69	0,76	0,56	0,53	0,59	0,69	0,47
Schooling	9,09	15,47	9,48	12,64	13,94	10,41	10,97	10,88	12,74	10,16
SOPORTE	207	108	305	104	57	81	12	256	67	452

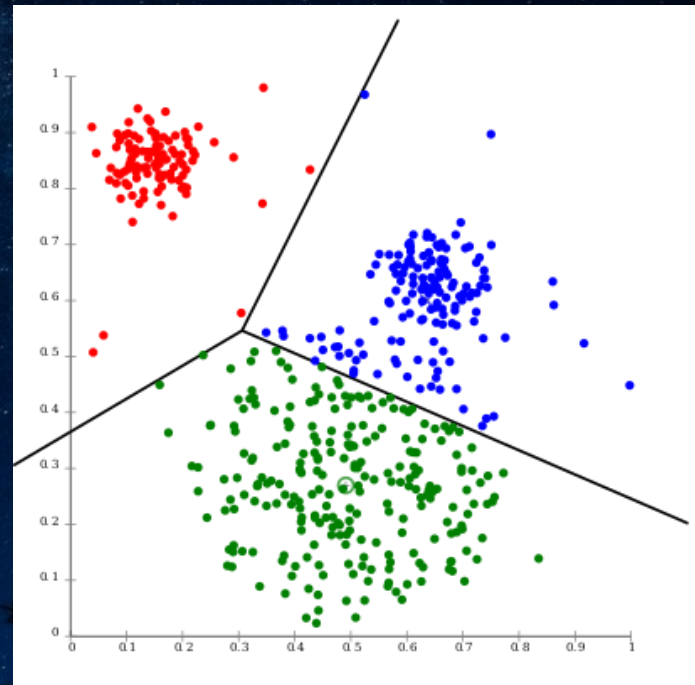


# Cuantización de datos con clustering

Los clusters permiten la eliminación de ruido en conjuntos con características similares.

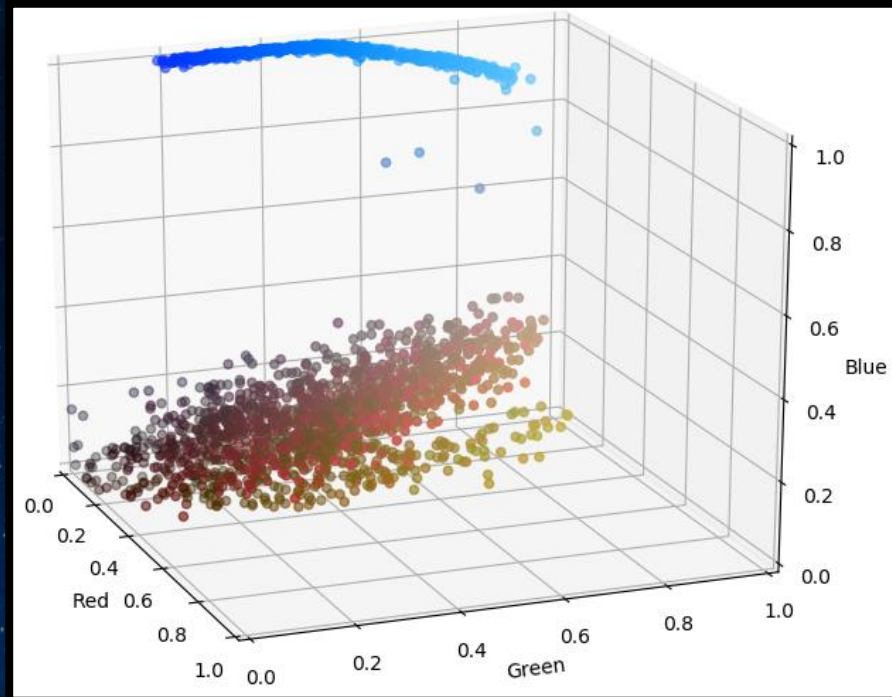
Esto podría utilizarse para reducir la cantidad de datos en un archivo (texto, imagen, audio).

La cuantización es una técnica de compresión con pérdida que consiste en agrupar un rango de valores en uno solo. De este modo, la codificación de los datos cambia, reduciendo el tamaño de almacenamiento.



# Cuantización de datos con clustering

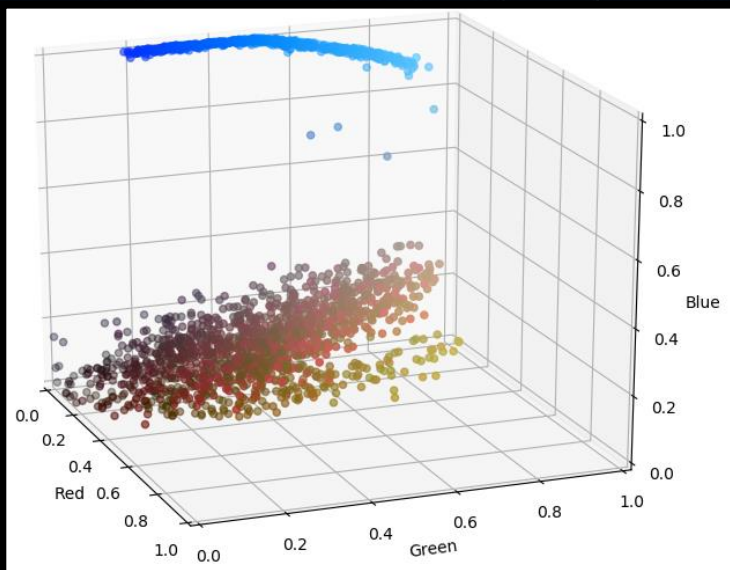
Supongamos una imagen RGB, y su distribución de píxeles.  
Cómo podemos reducir la cantidad de colores utilizando clustering?



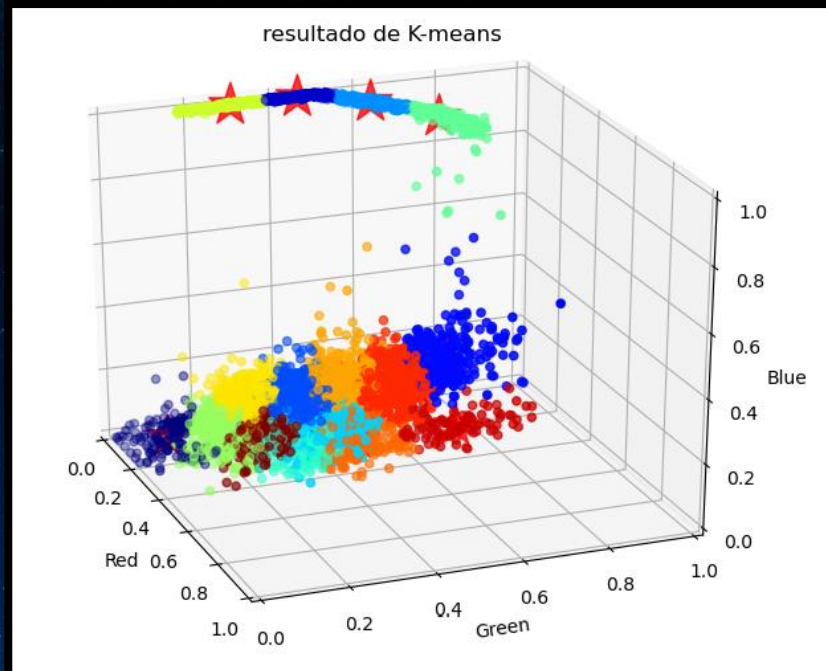


# Cuantización de datos con clustering

Distribución original



Asignación de clusters





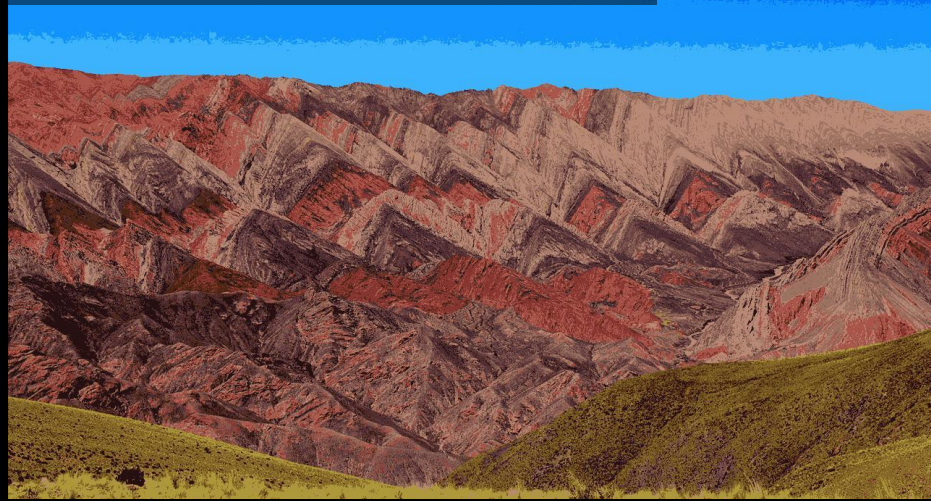
# Cuantización de datos con clustering

Imagen original 24bit de profundidad  
(16 millones de colores)



$720 \times 480 \times 24 \text{ bits} = 1,012 \text{ Kb}$

Imagen remuestreada a sólo 4 bits  
(16 colores)



$720 \times 480 \times 4 \text{ bits} = 168 \text{ Kb}$

# Elección de K

- El Error siempre decrece con K más grande
  - Peor caso:  $\text{Error}=0$  si  $K = \text{\#ejemplos}$
- Algunas técnicas:
  - Punto Elbow
  - Silhouette
- La elección de K casi siempre es subjetiva.  
**¡ No hay un K ideal !**

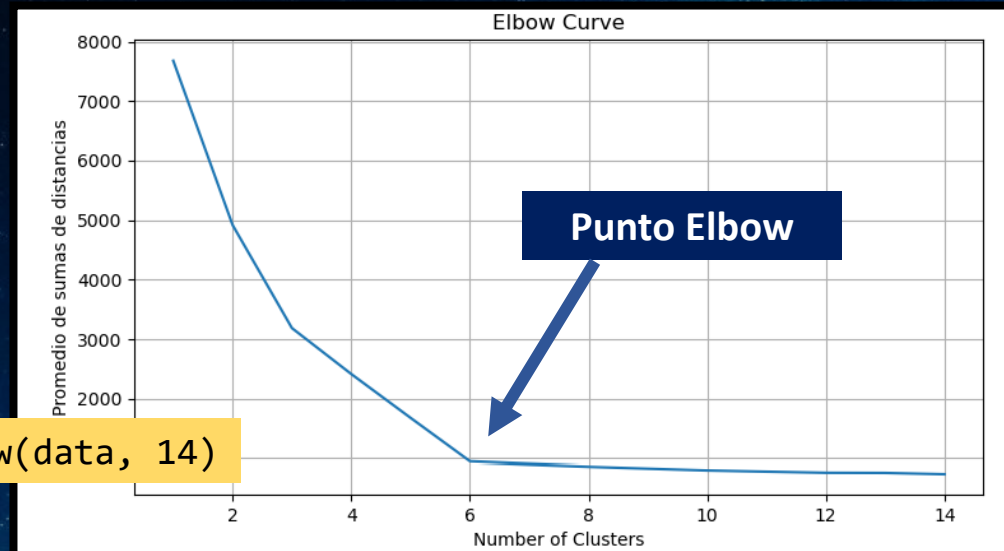


# Elección de K – Punto Elbow

**Punto Elbow** (punto codo).

Grafica el error del modelo para diferentes K.

El punto Elbow (si existe) se encuentra en el momento en que el error deja de bajar con la misma intensidad.



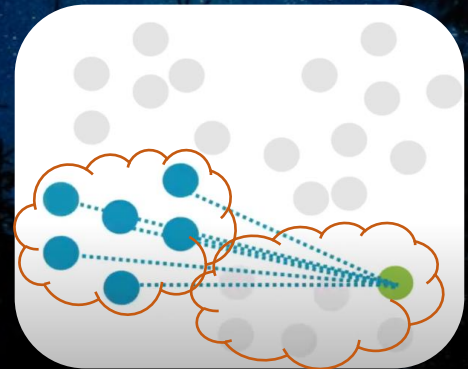
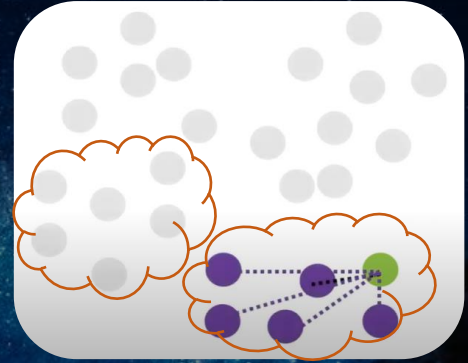
```
AA_utils.graficar_punto_elbow(data, 14)
```



# Elección de K – Índice Silhouette

## Mide cohesión y separación:

- Cohesión  $a(x)$  : distancia promedio de  $x$  a todos los elementos en el mismo cluster.
- Separación  $b(x)$  : distancia promedio de  $x$  a todos los elementos en el cluster más cercano.



# Índice Silhouette

**Índice Silhouette para un punto  $x$ :**

$$s(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

$a(x)$  = distancia promedio intraccluster

$b(x)$  = distancia promedio intercluster

**El valor de  $s(x)$  varía entre -1 y 1:**

1 = Buen agrupamiento

0 = Indiferente

-1 = Mal agrupamiento

# Índice Silhouette promedio

**Índice Silhouette para todo el clustering:**

$$SC = \frac{1}{N} \sum_{i=1}^N s(x_i)$$

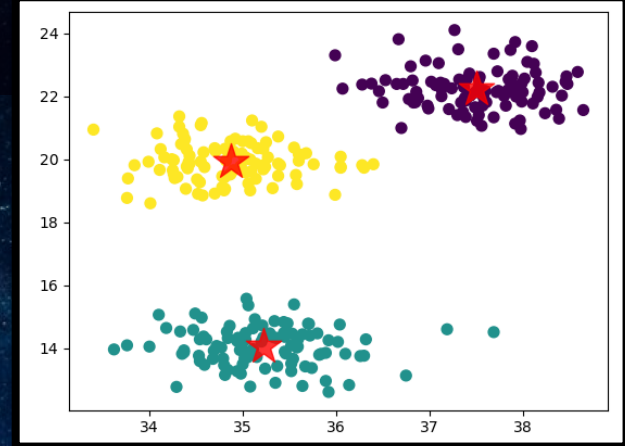
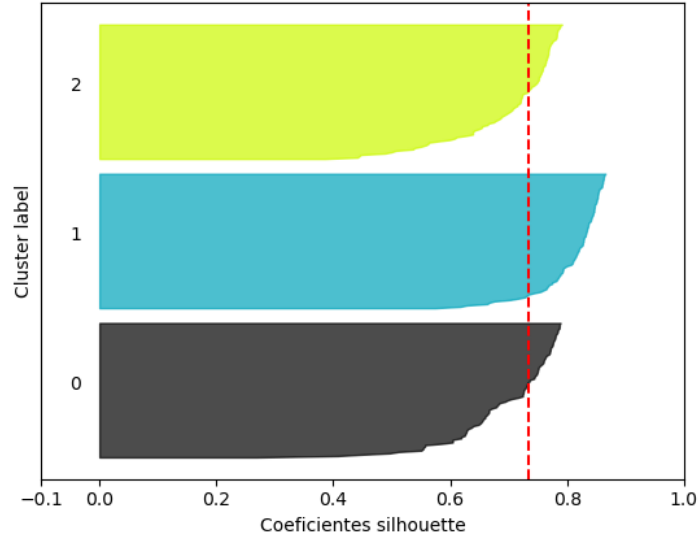
El índice el mejor cuanto mayor sea el número  
(más cercano a 1)



# Índice Silhouette

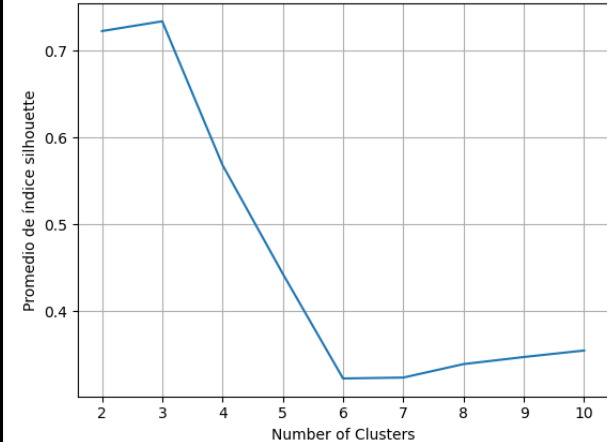
```
AA_utils.graficar_indice_silhouette_k(data, 3)
```

Silhouette plot para cada Cluster.



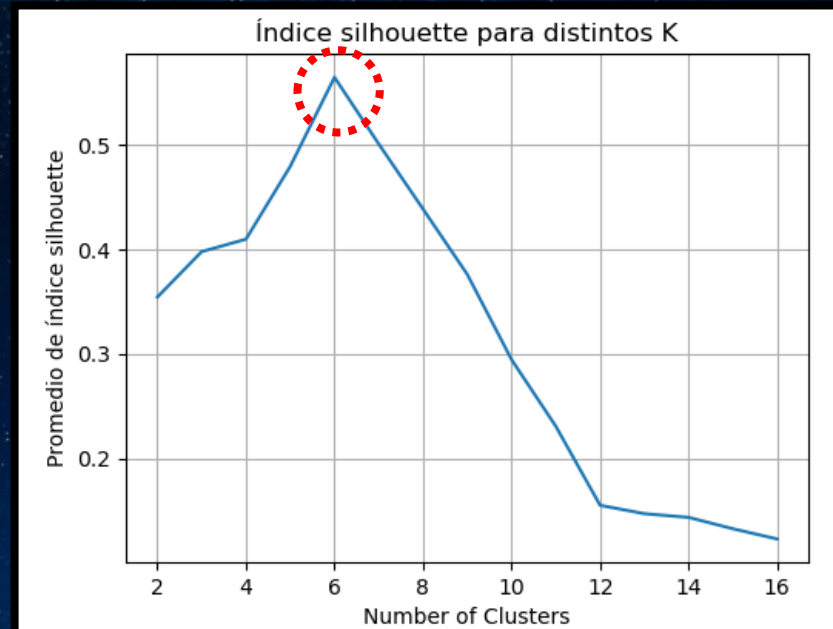
```
AA_utils.graficar_indice_silhouette(data, 10)
```

Índice silhouette para distintos K



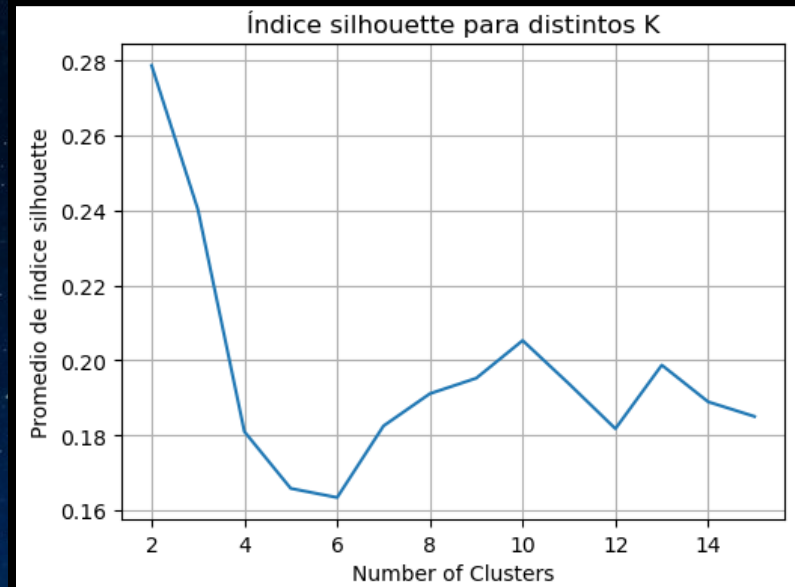
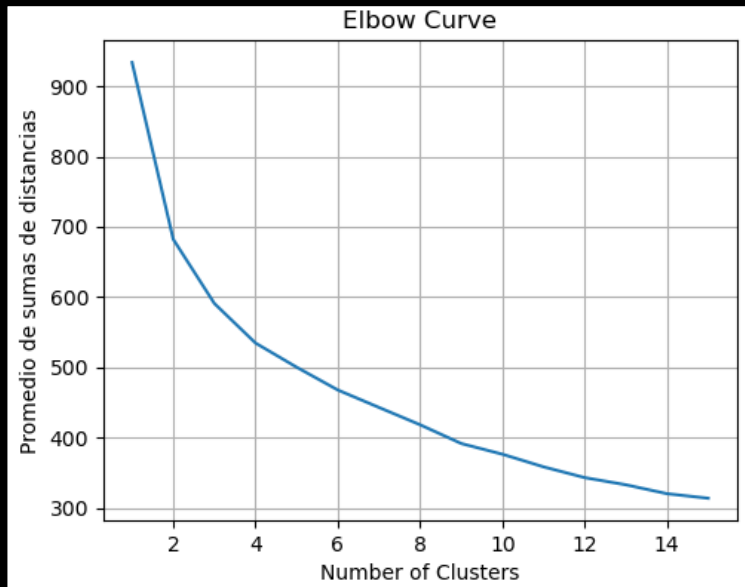
# Ejemplo - clustering\_10\_dim.csv

Vemos claramente que las curvas nos proponen elegir  $k=6$ .



# Ejemplo Países - WHO\_life\_expectancy.csv

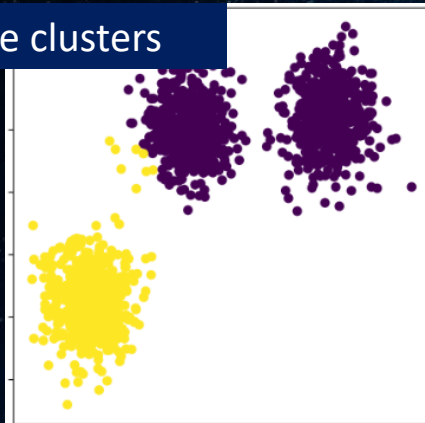
No podemos obtener un valor de K “óptimo”.  
Termina siendo una elección subjetiva.



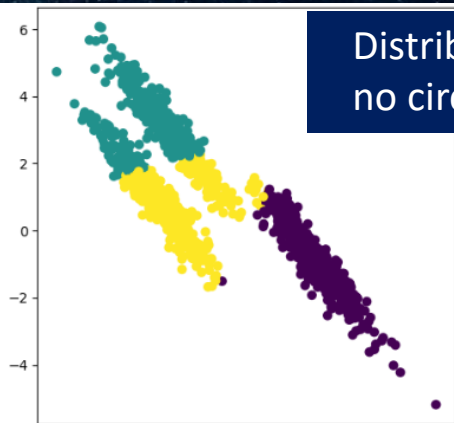


# K-medias. Resultados poco intuitivos

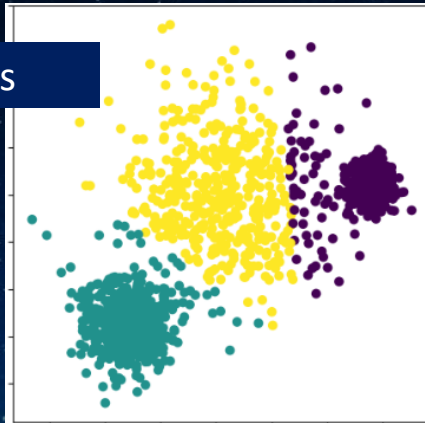
Incorrecto número de clusters



Distribución de datos desigual o no circular/esférica



Diferente varianzas



Clusters con pocos datos

