2A. Práctica: K-means

Nombre: **Luis Fernando Izquierdo Berdugo**

Materia: **Procesamiento de Información**

Fecha: **17 de Septiembre de 2024**

Instrucciones:

Leer el conjunto de datos del archivo `kmeans-elbow.npy` y aplicar la técnica del codo para estimar el mejor parámetro k del algoritmo k-means.

Se iniciará esta actividad importando los módulos de Python que se utlizarán, estos siendo `numpy`, `matplotlib` y de `Sci-kit learn` se importarán la función `KMeans`

In [160…
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

Se cargan los datos del archivo.

In [161…
```
X = np.load('kmeans-elbow.npy')
```

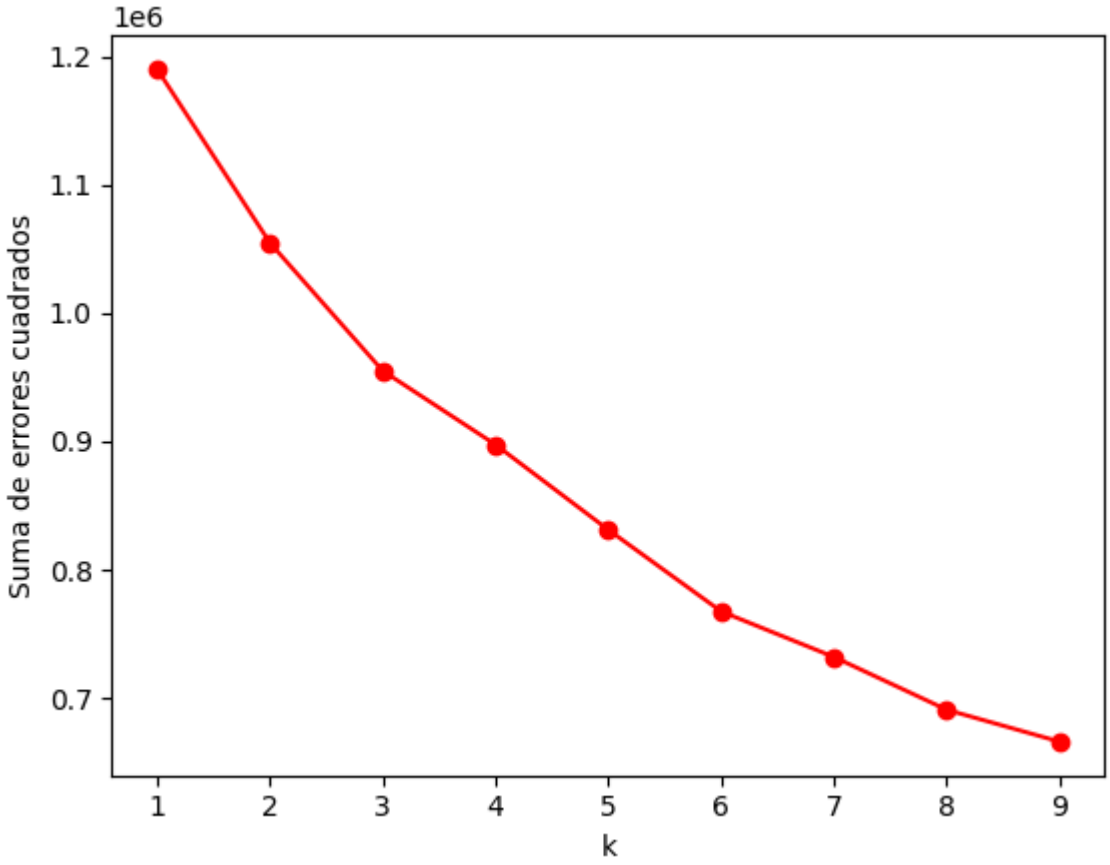
Se crea el código para la técnica del codo:

- Se crea la lista vacía `sse` que contendrá la suma de errores cuadrados (Sum of Squared Errors) para cada valor de K. Esto representa la distancia cuadrada total entre cada dato y su centroide asignado.
- Se crea un rango `k_range` que son la cantidad de clusters a evaluar.
- Se hace un ciclo que itera por todos los valores de `k_range`, el cual:
  - Crea una variable `km` que serán los KMeans creados, estos se crean usando la función `KMeans` con el número de clusters a crear, de igual manera se declara un estado de aleatoreidad para que el experimento sea constante al correrlo varias veces.
- Se usa el método `km.fit(X)` para que los datos de la variable `km` sean compatibles con los de la variable `X` (datos iniciales).
- Se accede al atributio `km.inertia_` que guarda los valores de suma de errores cuadrados y se añade a la lista `sse`.
- Se hace la gráfica del codo para visualizar cual es el valor óptimo de K

In [162…
```
def codo(rango):
    sse = []
    k_range = list(range(1, rango))

    for k in k_range:
        km = KMeans(n_clusters=k, n_init='auto', random_state=1004)
        km.fit(X)
```

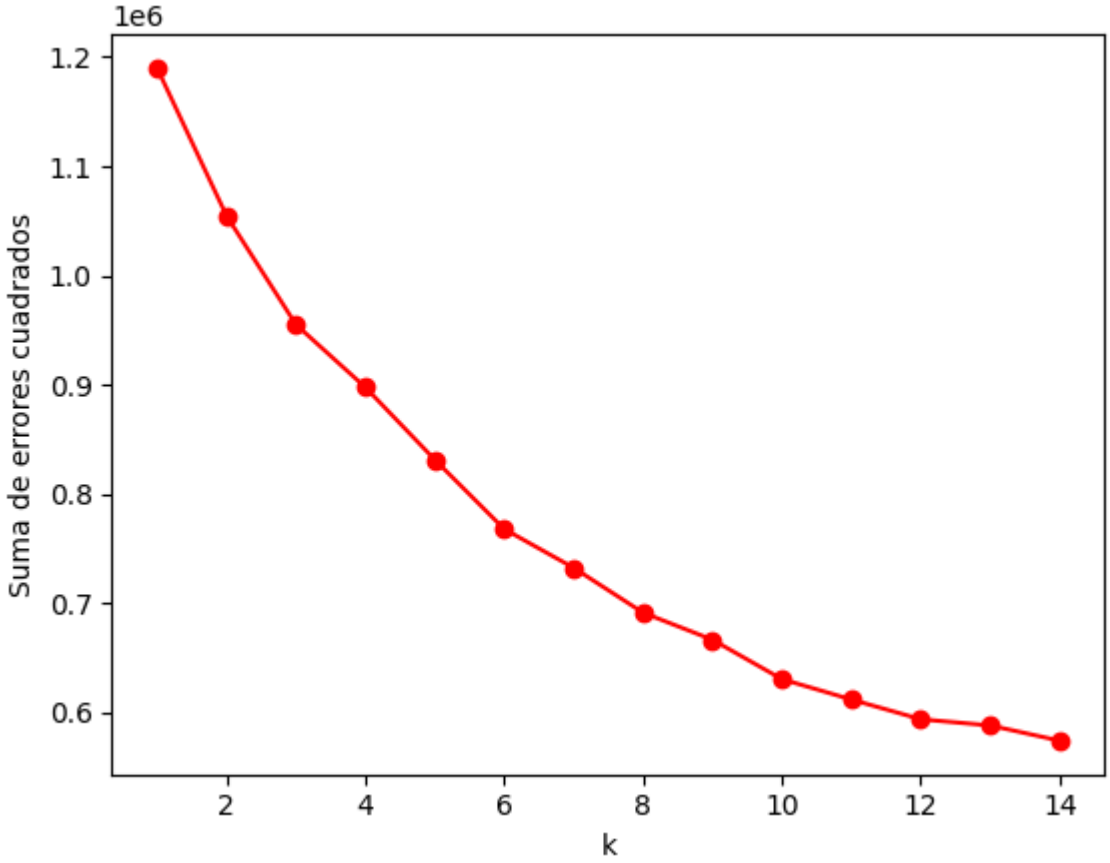
```
        sse.append(km.inertia_)

    plt.plot(k_range, sse, '-or')
    plt.xlabel("k")
    plt.ylabel('Suma de errores cuadrados')
    plt.show()
```

In [163…
```
codo(10)
```

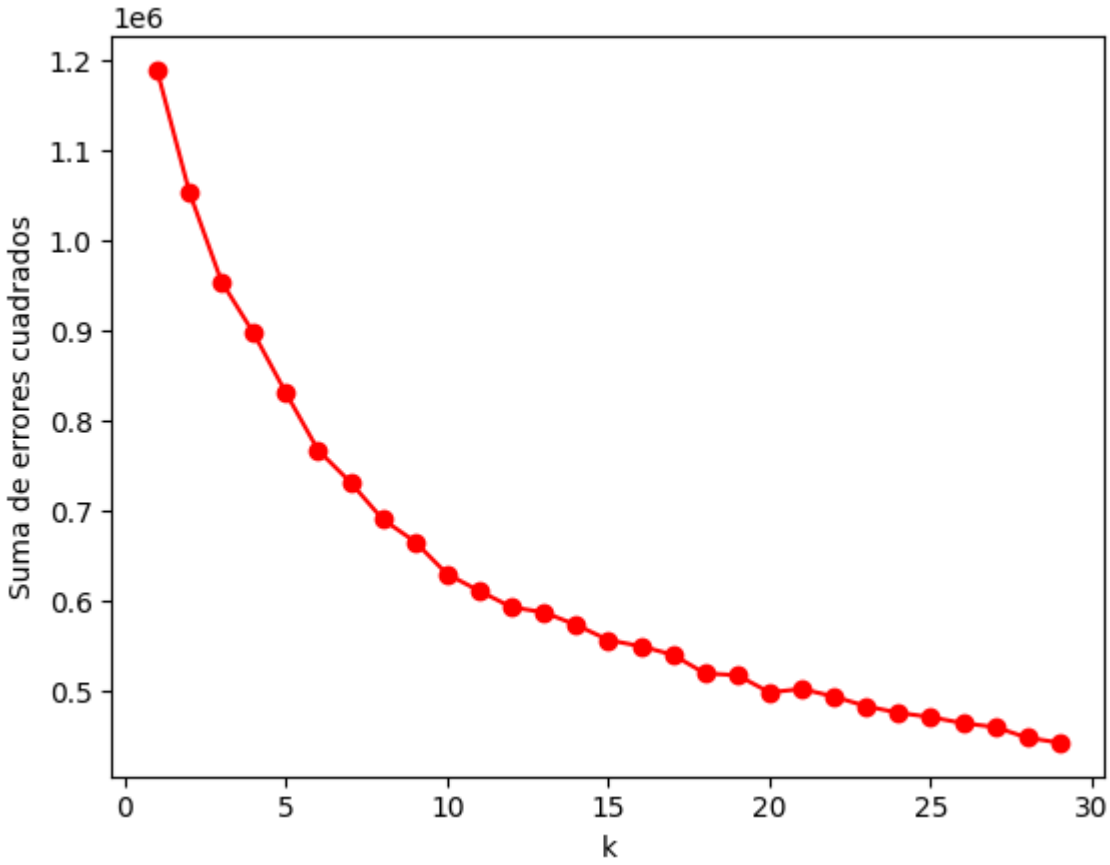


Debido a que no se ve un punto claro donde la curva se aplane de manera abrupta, se usan más valores de K a evaluar

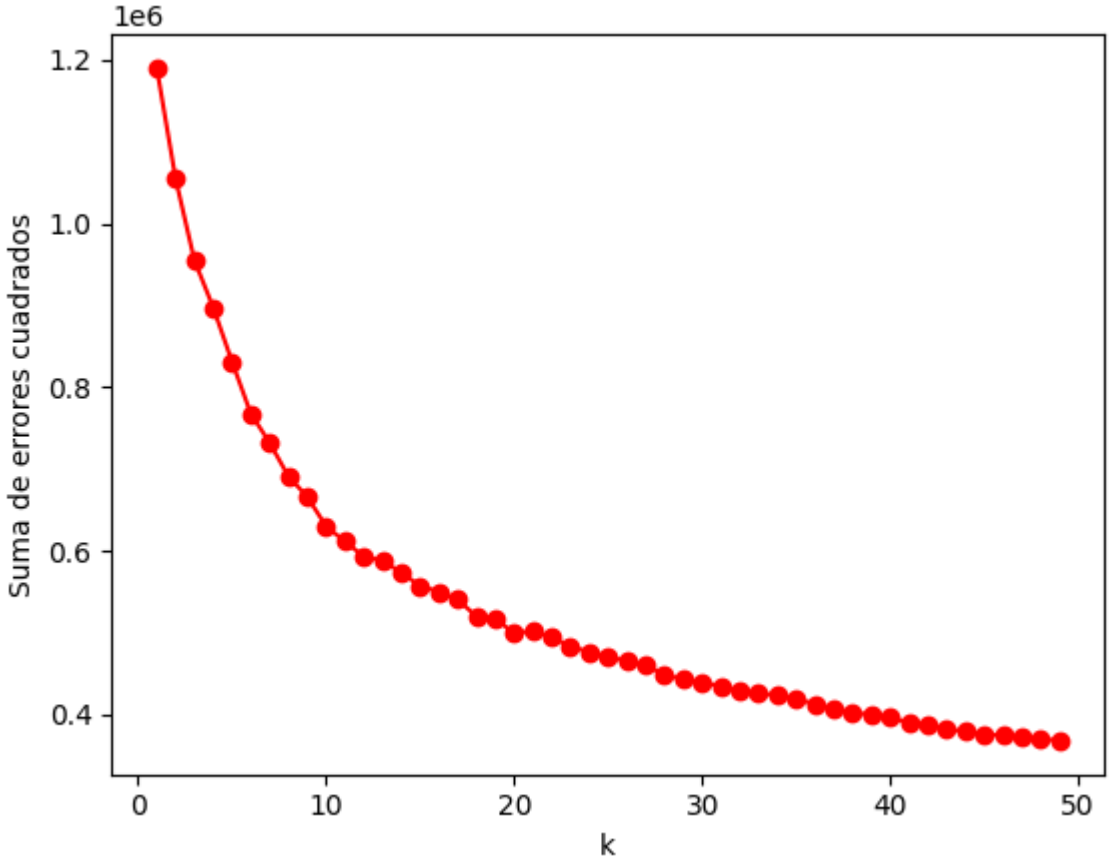
In [164…
```
codo(15)
```



In [165…
```
codo(30)
```



In [166…
```
codo(50)
```



De manera segura, no se puede usar un valor de `k` a simple vista. El único valor que tiene una significancia es el `8`, ya que podemos ver una disi¡minución en el espacio entre puntos posterior a este, lo cual podría indicar ser el valor a utilizar.