Clustering (unsupervised technique)

- grouping, similar to each other
- outlier detection (data cleaning / Processing)
- filling gaps in your data

partitional clustering

- n data, k clusters

$$\sum_{k}^{k \,\to\, k \text{ clusters}} \sum_{x_i, x_j \in \underset{\downarrow \text{ clustering}}{C_k}} d(x_i, x_j)$$

- Centroids → center of the cluster

$$\sum_{k}^{k} \sum_{x_i, x_j \in C_k} \overset{\text{euclidean distance}}{\overset{\uparrow}{\underline{d}}}(x_i, x_j)^2 = \sum_{k}^{k} |C_k| \sum_{x_i \in C_k} d(x_i, \underbrace{u_k}_{\text{mean of } C_k})^2$$

- always converge, but can't find the optimal solution every time, depend on the starting point - too close to each other

- use k-means ++ if possible