K - Means Clustering

- can be on line, graph, heatmap.
- using computer to identify / group data into meaningful clusters

K - Means clustering steps

1) Select number of clusters to identify (K)
2) Randomly select K data points to start (initial centroid)
3) Select selected data point & measure distance from all K centroids
4) Assign point to nearest centroid
5) Repeat until all points allocated

6) Calculate mean of each cluster
7) Reposition centroid to mean of each cluster
   ↳ reallocate points according to new centroid
   ↳ Stop once centroid converges.
8) Assess quality of clustering by measuring variance within clusters

↑

Repeat entire process with different starting points.

Determining ideal K

- As K increases, total variation ↓
- if $K = N$, total variation = 0

* consider reduction in variance for K

∟ Elbow plot to determine largest reduction in variance

Multi - Dimension K - Mean clustering

- Key idea behind K - Means clustering is finding Euclidean distance

Suppose 3 dimensions:

1) Randomly choose K centroids (x_1, y_1, z_1)
2) Calc. dist. b/w pt & centroid $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$
3) compute mean of cluster as new centroid : $\left(\frac{\sum X_i}{N_1}, \frac{\sum Y_i}{N_1}, \frac{\sum Z_i}{N_1}\right)$
4) Repeat 2 & 3 until centroid converge.
5) Evaluate Within - Cluster Sum of Squares (Inertia)

cluster | data points.

↳ used in Elbow method.