Exercise Clustering

https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset

→ The aim of this exercise is to get you stated on your first own clustering analysis. Use the r script provided on Monday by Mohanad and run it for the new dataset of your choice (diabetes or breast cancer). Afterwards, you can think about further adjustments, like using a different clustering method, different way of identifying the best number of clusters, different visualization methods.
Good luck!

Diabetes dataset

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

1. Number of Instances: 442
2. Number of Attributes: First 10 columns are numeric predictive values
3. Diabetes_label: Column 11 is a quantitative measure of "disease progression" one year after baseline
4. Attribute Information:
   - age age in years
   - sex
   - bmi body mass index
   - bp average blood pressure
   - s1 tc, total serum cholesterol
   - s2 ldl, low-density lipoproteins
   - s3 hdl, high-density lipoproteins
   - s4 tch, total cholesterol / HDL
   - s5 ltg, possibly log of serum triglycerides level
   - s6 glu, blood sugar level
5. Aim:
   - Use the attribute information to identify distinct clusters.
   - Compare "disease progression" (diabetes_label) across the clusters: look at mean and std

Breast cancer dataset

1. Number of Instances: 569
2. Number of Attributes: 30 numeric, predictive attributes and the class
3. Label: WDBC-Malignant = 1; WDBC-Benign = 0
4. Attribute Information:

   - radius (mean of distances from center to points on the perimeter)
   - texture (standard deviation of gray-scale values)
   - perimeter
   - area
   - smoothness (local variation in radius lengths)
   - compactness (perimeter^2 / area - 1.0)
   - concavity (severity of concave portions of the contour)
   - concave points (number of concave portions of the contour)
   - symmetry
   - fractal dimension (“coastline approximation” - 1)

5. Aim:

- Use the attribute information to identify distinct clusters.
- Compare the prevalence of breast cancer (computed via using the parameter “label”) across the clusters: look at mean and std