- Homework 5 Spec sheet -

In this homework, we are exploring a dataset on wines. This dataset is deliberately small, so you can try all methods we explored in the lecture and still finish running the code in a reasonable timeframe, even for the slower methods with a $O(n^2)$ complexity. As there are no labels, we have to use unsupervised machine learning methods to understand the space that wines inhabit. The learning goal is to realize how different dimension reduction methods yield similar and different results and how dimension reduction methods interface with clustering methods. So please load and use the "wines.csv" data file.
This dataset contains information of 178 of the finest wines.

The first row represents the column header.
Each row after that represents the information of one wine. There are 13 attributes that characterize each wine.

Columns represent (in order):

1) Alcohol content (in %)
2) Malic acid concentration (in g/L)
3) Ash content (in mg/L) [This is not as gross as it sounds, mostly minerals absorbed by the grapes through the soil]
4) Alkalinity of the Ash (g/L of potassium carbonate)
5) Magnesium (mg / L)
6) Total phenols (mg / L per epa method)
7) Flavonoids (mg / L)
8) Stilbenes (mg / L)
9) Proanthocyanins (mg / L)
10) Color intensity (reflecting opacity)
11) Hue (= color)
12) OD280 (Protein concentration)
13) Proline content (amino acid)

This data has been carefully curated, so there are no missing values. In this exercise, we will attempt to determine how many different kinds of wine there are. This might help with marketing and customer targeting.

Mission command approach: As per §4.5 of the Sittyba, we will tell you what to do ("answer these questions"), not how to do it. That is up to you. However, we want you to:

a) Do the homework yourself. Do not copy answers from someone else.
b) Restrict your methods (for now) to what was covered in the lecture/lab (in other words, PCA, MDS, tSNE, kMeans, Silhouette, kMedioids, EM, dBScan))
c) Include the following elements in your answer (so we can grade consistently):

Each answer should contain these elements:

1) A brief statement (~paragraph) of what exactly was done to answer the question (narratively explaining what you did in code to answer the question, at a high level).
2) A brief statement (~paragraph) as to what you were thinking - some kind of rationale as to why made specific choices in your modeling, if you did make them.
3) A brief statement (~paragraph) as to what was found. This should be as objective and specific as possible – just the results/facts. Do make sure to include numbers and a figure (=a graph or plot) in your statement, to substantiate and illustrate it, respectively.
   As the unsupervised methods often yield visualizable results, be sure to include a figure.
4) A brief statement (~paragraph) as to what you think the findings mean. This is your interpretation of your findings and should answer the original question.

Note: Brief actually means "brief". There is no need to write a novel. There is value to being concise. A few pages should be sufficient for the entire report. Do – however – write a report. A data and code-dump is not very useful or valuable in practice. People who pay you so they can ask you questions usually want them answered. Succinctly. Clear, concise and coherent answers are valuable.

Please answer the following questions in your report:

1. Do a PCA on the data. How many Eigenvalues are above 1? Plotting the 2D solution (projecting the data on the first 2 principal components), how much of the variance is explained by these two dimensions, and how would you interpret them?

2. Use t-SNE on the data. How does KL-divergence depend on Perplexity (vary Perplexity from 5 to 150)? Make sure to plot this relationship. Also, show a plot of the 2D component with a Perplexity of 20.

3. Use MDS on the data. Try a 2-dimensional embedding. What is the resulting stress of this embedding? Also, plot this solution and comment on how it compares to t-SNE.

4. Building on one of the dimensionality reduction methods above that yielded a 2D solution (1-3, your choice), use the Silhouette method to determine the optimal number of clusters and then use kMeans with that number (k) to produce a plot that represents each wine as a dot in a 2D space in the color of its cluster. What is the total sum of the distance of all points to their respective clusters centers, of this solution?

5. Building on one of the dimensionality reduction methods above that yielded a 2D solution (1-3, your choice), use dBScan to produce a plot that represents each wine as a dot in a 2D space in the color of its cluster. Make sure to suitably pick the radius of the perimeter ("epsilon") and the minimal number of points within the perimeter to form a cluster ("minPoints") and comment on your choice of these two hyperparameters.

Extra credit:

a) Given your answers to all of these questions taken together, how many different kinds of wine do you think there are and how do they differ?
b) Is there anything of interest you learned about wines from exploring this dataset with unsupervised machine learning method that is worth noting and not already covered in the questions above?