Assignment 4: Advanced Topics (10 Points)

General instructions.

1. Restate your chosen SDG in your submission and complete all tasks by visualizing phenomena that speak to your chosen SDG.
2. You can reuse the dataset from the previous assignment or use one or more other datasets, provided that they are all connected to your SDG and contain the information needed to perform the assignment tasks.
3. Always specify which datasets you used for which tasks and provide links to the sources of all datasets.
4. Always specify which visualization tools (e.g., programming languages and libraries) you used for which tasks.
5. You are allowed to use generative AI *to support your learning*. If you opt to use generative AI, you *must* state which model(s) you used for which tasks, describe how and why you used them, and provide a critical reflection on how they supported you in understanding the course material and carrying out your assignment tasks. Undisclosed usage of generative AI, if discovered, will be considered cheating.

If you have questions or feedback on the assignments, please share these in the course forum.
See also our MyCourses page for further general instructions.

Specific instructions for Assignment 4 To complete each of the following tasks, please

1. create the visualizations as described in the task,
2. describe and interpret the visualizations with a view to understanding the progress toward your SDG (for Task 1) or all SDGs (for Task 2),
3. **discuss the impact of methodological choices** (i.e., which method, data preprocessing, initialization, and hyperparameters) **on the perception and interpretation of the relationships between your observations**, and
4. report any challenges you may have encountered and how you overcame them.

Your answers to 2.–4. should be at least one paragraph of text each. The target length of your text (excluding graphics) is between two-thirds of a page and one page, but you can remain below that if your answers are concise or go over that if you have more to share (no need to optimize the layout to make your answers look longer or shorter).

Task 1: Dimensionality Reduction (5 Points)

Identify a dataset pertaining to your SDG that has at least 10 observations (rows) in four distinct variables (columns). (You can probably reuse your dataset from Assignment 2, Task 3.) Use PCA, MDS, and t-SNE to reduce the dimensionality of your data and visualize it in two dimensions. Specifically, visualize

1. PCA once without standardization and once with standardization (2 figures in total);
2. MDS on the non-standardized data with two different random seeds (2 figures in total); and
3. t-SNE on the non-standardized data with two different random seeds for each of three different values of the "perplexity" parameter (6 figures in total).

Record the specific configurations you used to produce each figure. You can choose how to annotate and style the markers (e.g., color, shape) based on any (meta)data available to you, but make sure that the individual observations can be readily identified.

Task 2: Relational Data (5 Points)

Revisit the list of all Sustainable Development Goals and create a network dataset recording the interconnections that you personally see between these goals: Each SDG is a node (so 17 nodes in total), and undirected edges record your perceived interconnections (you can assign them weights or leave them unweighted). You will probably see many interconnections, but make sure that each goal is connected to at least one other goal and that you have at least 25 and at most 50 edges.[1]

[1] The easiest way to create a network dataset based on an edge list is to create a CSV file with one header row (e.g., "node1,node2") and then one row per edge (e.g., "1,2" for a connection between SDG 1 and SDG 2). This file can then be read with a dataframe library (e.g., pandas in python) and used to construct a network using a network library (e.g., networkx in python).

Write one paragraph describing your data-modeling process and the reasoning that went into your edge definitions, and also include your edge list as a table in your submission.
Then visualize the resulting dataset using

1. a radial layout with numerical node ordering (1 figure in total),
2. a Kamada-Kawai layout with two different random seeds (2 figures in total), and
3. a Fruchterman-Reingold layout with two different random seeds (2 figures in total).

Record the specific configurations you used to produce each figure. You can choose how to annotate and style the nodes (e.g., color, shape) and edges (e.g., color, weight) based on any (meta)data available to you, but make sure that the individual observations can be readily identified.