# CS 470: Data Mining

Homework I - Data Exploration and Preprocessing

Prof. Kai Shu
Due at Feb. 10th 2025, 11:59 PM

**Submission Instructions**: Submit your assignment through the Canvas system. Upload a single ZIP archive file named $hw1\_ < yourfirstname > \_ < yourlastname > .zip$, containing the report as well as three .csv files from Task 2 (see the details below). No email submissions are accepted. At the top of your solution, include a section named "Collaboration statement" in which you acknowledge any collaboration, help, or resource you used or consulted to complete this assignment. You can use any programming language(s) to perform the computations and visualizations.

*Note: Please observe the Emory University Honor Code. An automatic plagiarism checker will be used on submissions.*

The file **data.csv** contains a dataset of the information of anonymized patients' hearts. It contains 15 attributes. Perform all the data exploration and preprocessing tasks described below.

- [**Task 1 (20 points): Attribute Description**] Describe all the attributes of dataset in the report. For each attribute, explain its meaning (understanding the concepts of the attributes may help), and determine the type of attribute (nominal, ordinal, numeric interval, numeric ratio) – be careful that some of these attributes may be deceiving! If there is a "grey area" explain all the possibilities and always motivate your choices.

- [**Task 2 (35 points): Proximity Calculation**] Based on the material covered in class, please summarize the differences between proximity measures (correlation, cosine similarity, and Euclidean distance) (8 points). Then, calculate these **three** measures among all the pairs of data instances in the table for each measure. Please do it in two steps: First, remove all the non-numerical attributes; Second, calculate the proximity for all the pairs of data instances, show the formulas you used in the report, and store the results in a .csv file for each measure. In the report, make sure to provide enough descriptions of this .csv file to make it understandable, such as the meanings of the columns and rows.

- [**Task 3 (15 points): Summary Statistics**] For each attribute where it makes sense to do so, calculate the most common summary statistics: mean, standard deviation, and 5-number summary. Present your results in the report.

- [**Task 4 (20 points): Charts**] Generate 10 or more plots from the data: at least 3 box plots, at least 3 histograms, and at least 3 scatter plots. Include all the plots in the report. In the report, for each plot write a detailed description and a discussion of what can be observed and noticed about the data.

- [**Task 5 (10 points): Tools and Languages**] In the report, describe, compare, and discuss the pros and cons of all the various tools and languages that you used (or tried out) to complete the tasks in this homework assignment.

### Grading Criteria

- You will get full score for each task that is complete, correct, and well justified; you will receive a partial score proportional to the level of completion and correctness if the task is partially incomplete, or with minor mistakes, or not well justified; zero point if the task has major omissions, major errors, or it has wrong or totally missing justification.
- -10 points for missing collaboration statement in the situation when you should have it.