# CS 5402 – Intro to Data Mining
# Fall 2019
# HW #7

- This assignment is **due in class on Wednesday, Nov. 13, 2019**.
- This assignment is worth **100 points**.
- You are **REQUIRED to work as part of a team of 2-3 people**, each of whom must be a person enrolled in this course. That includes distance ed students!

## Project Description

For this assignment you are to use techniques discussed earlier this semester to **preprocess** a large (…well, not tiny) dataset that is posted on Canvas along with this assignment; we would prefer that you **ONLY** use techniques/methods that were discussed in this class. You will **NOT** be given **specific instructions** about what procedures to perform on the dataset; your grade on this assignment will be based on your determination of **what preprocessing should be done, how to do those tasks, and the best (i.e., most efficient and logical) order in which to do those tasks.**

You can use any software applications and/or programming languages to do the preprocessing. For every preprocessing task that you do, you must **provide brief technical documentation** explaining **how** you did it. For example, if you used a particular filter in Weka, include a screenshot of the KnowledgeFlow program that you set up to perform that task. If you wrote a program to do some task, include the source code for that program. If you did a sequence of actions in Excel, just briefly explain what you did. Do **NOT** submit lengthy/verbose explanations of what you did – we won't read (or grade) more than the first 3 sentences you write! Even with people working in teams, the grader has to grade numerous reports in a relatively short amount of time ☹ If you do not provide **concise, precise documentation for each method** you used, you will not receive full credit!

The first 1-2 pages of your homework submission should simply be a **table** containing about 3 columns: (1) a column (briefly) saying **what** you did, (2) a column saying **how** you did it, and (3) a column (briefly) saying what the **purpose** was for doing that task. You also could include a fourth column that gives a reference to another page in your report where the source code, KnowledgeFlow screenshot, etc. can be found that provides additional (more verbose) documentation for the method used. These entries should be listed in the table in the same order that you actually performed them. **Note that you will lose points if you do not do things in a reasonably efficient/logical order!**

## What To Submit for Grading

You should submit a **paper** copy of your report in class the day it is due. Do **NOT** turn in a printed copy of the data file! **The grader reserves the right to contact you and ask to see a digital copy of your processed data file.** It is your responsibility to have that file available to show him upon demand (i.e., if he contacts you, you can't say "the system ate my file and I don't have it anymore"); if you don't have the file to show him when he asks for it, you will get a zero on this assignment! You should have every member of your team make a backup of the file!

You are **also required to submit ONLINE (via Canvas) a survey/evaluation of your team members**. It will ask what tasks you and each member of your team did on this assignment, and what percentage of credit (e.g., full vs. partial) you think each team member deserves. This will be taken into consideration when determining your grade on this assignment. If the tasks/credit that you claim for yourself vastly differ from what your team members state for you, then a meeting will be held with your instructor (Dr. Leopold) and possibly the Computer Science Department Chair (Dr. McMillin) to determine if academic dishonesty has taken place; so be honest! **If you do not complete the online survey/evaluation, you will receive zero on this assignment, even if you worked on the project.** The survey will be posted on Canvas as an online "quiz" called **HW #7 Evaluation**.

**Hint:** To receive full credit on this assignment you should consider preprocessing ideas that were discussed **throughout** the semester, not just the things that were discussed in the lecture titled "Data Preprocessing & Warehousing."