CMS 495 – CS Topic: Machine Learning
Assignment: Machine Learning Data Cleaner -- Dr. Jennifer Seitzer

PART 1: Data Generator – Due in one week

Consider the data presented below. Create a file of fifteen records that abide the following file schemea and then represent the file as a CSV (Comma Separated Value) file. Your file will be combined with the other groups' files to form one large data set on which we will apply Part II (below). You should also create a file to be used in the "JOINING" phase as indicated in the In-Class Handout. **Thus, submit two CSV files of fifteen records each.**

Client Number	Name	Address	Date of Purchase	Magazine Type
23003	Johnson	1 Downing St.	04-15-1999	Car
23003	Johnson	1 Downing St.	06-21-1999	Music
23003	Johnson	1 Downing St.	05-30-1999	Comic
23009	Clint	2 Boulevarde	11-11-1111	Comic
23013	King	3 High St.	02-30-2000	Sports
23019	Jonson	1 Downing St.	11-11-1111	house

PART 1I: Program a Data Cleaner

Build a system to transform the data, one step at a time, following the sequence of steps covered in class as the KDD (Knowledge Discovery Process) for the Magazine Marketing Problem, clean the data. The steps to be applied include:

-- Selection (from part I above)
-- De-duplication
-- Correction
-- Joining
-- Coding :
  --address to region,
  --date to period,
  --income divided by 1000,
  --binary transformations

1. You may implement each of the above (sans "Selection") as its own class or method
2 You may build this as a Menu-Driven Program or with a GUI
3. You will demo your system in two weeks to the class. The final collective data file of "dirty data" will be available in one week and one day. So test your system on your own file until the full file is available.