Mini Project #2

Practice the three basic tasks of visual data analytics

- use data from mini project #1 (or other), begin with |N|≥500, |D|≥10)
- client-server system: python for processing (server), D3 for VIS (client)

Task 1: data clustering and decimation (30 points)

- implement random sampling and stratified sampling
- the latter includes the need for k-means clustering (optimize k using elbow)

Task 2: dimension reduction (use decimated data) (30 points)

- find the intrinsic dimensionality of the data using PCA
- produce scree plot visualization and mark the intrinsic dimensionality
- NEW: show the scree plots before/after sampling to assess the bias introduced
  you could also visualize the before/after sampling data via MDS (see below)
- obtain the three attributes with highest PCA loadings

Task 3: visualization (use dimension reduced data) (40 points)

- visualize the data projected into the top two PCA vectors via 2D scatterplot
- visualize the data via MDS (Euclidian & correlation distance) in 2D scatterplots
- visualize scatterplot matrix of the three highest PCA loaded attributes

NEW due date: Tuesday, 3/12

Note: this is a time-intensive project – start early!!

Deliverables

You need to upload to Blackboard the following by the due date:

- 2-3 page report with illustrated description of your program's capabilities and implementation detail
  - add code snippets to show how you did things
  - discuss interesting observations you made in the data
  - constructively compare the various alternatives in task 1,2,3
  - make good use of visualizations
  - video file that shows all features of your software in action
  - archive file with source code

Grading

- TA will pick students at random for thorough code review sessions
- you better know your code !!!
- so, please do not just copy code beyond the D3 templates
- or even worse, videotape someone else's program