Too many data items?
  - k-means clustering followed by stratified sampling

Too many attributes?
  - PCA followed by dimension projection and scree plot culling

Not enough data items or attributes?
  - data synthesis using data augmentation

Use Python library scikit-learn
  - lots of simple and efficient tools for data mining and data analysis
  - is available on the department’s Linux server via ssh login into nlclient22.cs.stonybrook.edu (port 130)
  - use putty to login and upload

You will earn 10 pts extra credit for whatever option you pursue
  - 10 pts are regular credit, additional options earn 10 pts extra credit each
Visualize the data you secured in project #1 using D3 bar charts

Your D3-based visual interface should be able to (all 10 pts):
1. pick a variable and bin it into a fixed range (equi-width) of your choice
2. create a bar chart of the variable you picked in 1.
3. respond to mouse clicks to cycle through all variables and update chart
4. only on mouse-over, display the value of the bar on top of the bar
5. on mouse-over, make the bar wider and higher to focus on it
6. create a drop-down menu to allow users to select the variable shown

An additional 10 pts for elegant implementation/function

Extra credit (10 pts):
- mouse moves left (right) with left mouse button depressed should decrease (increase) bin width/size

Submission via conference website
Due Monday, Oct. 15, 2018
You need to upload the following by the due date

- please use Blackboard for your submission
- 2-3 page report with illustrated description of your program’s capabilities and implementation detail
- data processing (2.1, if you did it) and visualization (2.2)
- your can add code snippets as an appendix
- video file that shows your software in action
- as before, do not reveal your identity on your submission

Also upload your source code by the due date

- zip it and upload it to the submission site (choose project 2)
- do not upload it to Blackboard because it would be shared with your peers

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
Video recording

- a good program is **Apowersoft Screen Recorder**
- captures screen and voice at the same time
- it’s free for a version with sufficient capabilities
Aka, cheating

Discussion with your class mates is OK
- stackoverflow.com is a good source of information
- to get some advice on implementation details when you get stuck
- not meant to be a shortcut to avoid own implementation work

Cut and paste from any source is not OK
- any suspected activity of this kind will result in zero points
- also for the person providing the original
- two-strikes and out rule is in effect (including an academic misconduct report)
- this includes any feeble attempt to cover the tracks somehow

Stay honest and resist the temptation!
You will all be assigned 3 random peer project reports
- you will serve as a reviewer
- we will send email on specific login procedures and deadlines

Two review criteria:
- Does the report describe the chosen data and their domain well?
- Does the report describe the prospects of the data for interesting insight well?

Review will be along these two dimensions:
- scoring with the Likert scale
- verbal comments