NAME : SRAVANI KAMISETTY
SID : 304414410
COURSE : CS170 – MATHEMATICAL MODELLING AND METHODS
DETAILS : PROJECT PROPOSAL
TOPIC : MORSE CODE: INSIDE THE COLLEGE RANKINGS

Dataset
I would like to work with the dataset from 1995 US News and World's Report Guide to America's Best colleges. I found this dataset on the CMU's StatLib website. The U.S. News data contains information on tuition, room & board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio and a number of other variables for 1300+ schools. Also, there is a related dataset- the AAUP. It includes average salary, overall compensation, and number of faculty broken down by full, associate, and assistant professor ranks.

Why this dataset?
I am particularly interested in this dataset because I feel that I picked certain colleges over others when I applied to my Masters in America. I would like to know whether the factors that influenced my decision are the same factors that grade/rate the colleges.

What can be done with this data?
I did a little bit of exploratory data analysis on the data and found that these are some of the many things which can be done with this data.

1. Model tuition using the other variables
2. Clustering colleges into similar comparison groups
3. Best way of structuring faculty salary
4. Reasonable way to rank the schools
5. Try to predict the type of school i.e either public or private (may be by using a classification procedure)
6. Finding relationship between college dataset and professor dataset

During the data analysis, I found that the dataset has a few missing values. In the project I would like to use a data mining procedure (like nearest neighbours) to fill up the missing values. Also, PCA could be done after and before the prediction to see the changes, if any in the principal components.

I feel there is lot more information locked in this dataset. Exploratory factor analysis can be performed to learn the latent variables underlying in the data that could be related, but not observed. This could help in the categorization of the data. It would be very interesting to compare the factors that affected my decision in picking colleges with the latent variables.

Location of the dataset
American Statistical Association, 1995 Data Analysis Exposition. US News and World Report's Guide to America's Best Colleges, 1995. available on Statlib website: http://lib.stat.cmu.edu/datasets/.