I was interested in looking at patterns among a small set of universities to see if any interesting patterns existed. The reason the dataset ended up being small (21 Universities) was because the data had to be collected by hand. Some universities keep their records in what is known as the Common Data Set Format. A handful of these universities publish these yearly reports on their website, ranging in formats from PDF to HTML to XLS. It was pretty easy to find the forms by using google because the format remains the same within each document. The form consists of 10 sections (General Information, Enrollment, First-year Admissions, Transfer Admissions, Academic Offerings, Student Life, Annual Expense, Financial Aid, Faculty and Class Size, and Degrees Conferred). I only extracted sections from General Information, Enrollment, First Year Admissions and Annual Expense (Tuition).

The dataset consists of 21 records with 27 attributes each. The attributes collected include:

Name, Funding, Coed Or Not, Calendar, Full Time Students, % Men, % Women, % Aliens, % Black, % Indian, % Asian, % Hispanic, % White, % Unknown Race, Men Applied, Women Applied, % Men Admitted, % Women Admitted, % Men Enrolled, % Women Enrolled, SAT Verbal 25th, SAT Verbal 75th, SAT Math 25th, SAT Math 75th, ACT 25th, ACT 75th, Tuition

Some of these attributes were calculated and/or cleaned up before importing them into HCE. For instance, none of the attributes were found in percentages, but to compare attributes between universities, you have to normalize the values against the total value of that university. HCE does some normalization, but it cannot figure out how to do advanced normalizations like this, so all of these had to be done in Excel as a preprocessing step.

Goals

The main goal of our analysis was to find groups of similar universities and to find patterns and trends between the attributes. Clustering is the ideal choice for the first task and so HCE was used. HCE also provides a good way to use 1D and 2D histograms for discovery. HCE proved to be rather straightforward and easy to use. Importing files was very easy and it did not take long to start finding obvious clusters. Cycling through the histograms ended up taking a while, especially since many of them were such obvious relationships that it would have been nice to be able to mark them in some way such that they would be skipped (e.g. SAT Verbal 25th vs. SAT Verbal 75th).
Results

For obvious reasons the attributes to cluster on did not include whether a school was public or private and also did not include whether it was an all womens school or not. This was done because I thought they would make good natural clusters that hopefully HCE could pick up on without knowing the classes for sure. This was in fact the case as is shown in Figure 1 and Figure 2.

Figure 1. The first cluster shows the obvious separation between the public and private schools. The Cluster that is selected consists of all the public schools and no private schools.

The public schools selected here were University of Maryland, University of Florida, Georgia Institute of Technology, University of Connecticut, University of Michigan Ann Arbor, Umass Amherst, and University of Pittsburgh.
The private schools in the dataset include: Amherst College, Barnard College, Cornell University, Dartmouth College, MIT, Mount Holyoke College, Northwestern University, Oberlin College, Princeton University, Smith College, Stanford University, Swarthmore College, Williams College, and Yale University

Figure 2. HCE also correctly clustered the three women's schools (Smith, Mount Holyoke, and Barnard) correctly into their own cluster which is highlighted here.
Figure 3. The Tuition Gap

The tuition graph between public (red and leftmost blue) and public (the rest of the blue) schools can be seen clearly in this outlier graph produced by HCE. The leftmost blue is University of Michigan Ann Arbor. This was presented as the largest gap in the dataset which should probably not be a big surprise. One thing to note is that the out of state tuition price was used for all public schools as to not to bias the results too much, however, the difference is still enormous.

Another interesting graph presented by HCE was the comparison of the number of applications by men verse the number of applications by women. Overall this is a very linear relationship as would be expected. The obvious exception to this case is the three womens schools which are clearly found along the X axes in Figure 4. The other outliers are somewhat more interesting, they are the schools where more men applied then women. These two schools were MIT and Georgia Tech, the two technology schools in the dataset.
Another interesting relationship that was presented by HCE was the one between the percent of women being admitted to the percent of a student body being white. Of course these sorts of conclusions could start to walk down a slippery slope, so I will not make any personal reflections about the possibilities why this relationship could exist. It does however appear that at a university where there is a greater white population that a female will have better chances of getting in. A similar relationship existed for men as well, but the slope was not as great. It’s important to note that the schools in the upper right of the graph are mostly larger state schools, where acceptance rates are expected to be higher. See Figure 5 for more details.
Figure 5. Relationship between the diversity of a school and the chances of a woman being accepted.

Figure 6. Schools with higher average SAT scores are harder to get in.
The relationship shown in Figure 6 is a rather obvious one; however, it is interesting to see some of the outliers in this case. For instance, one of the obvious outliers (when viewed in HCE with the linear correlation line overlaid on the histogram) is Oberlin. Of course we all know that Oberlin is incredibly difficult to get into, however it is not strongly based on SAT scores but rather musical talent.

Figure 7. Getting the biggest bang for the buck.

Most people would consider that the harder a school is to get into then the more likely it is a very good school for one reason or another. A good way to find out which school is the best for the amount that they want in return for tuition is to look at a histogram comparing the tuition verse the admittance rate. As can be clearly seen here, the data point marked by the triangle is most likely the “best bang for the buck” in terms of schools. As should be no surprise this data point represents the University of Maryland.

Final Thoughts

Overall I found working with this dataset was not only interesting but also fun as well since the relationships were obvious and easy to draw conclusions from. I thought that using HCE was pretty straightforward and it proved to be an extremely powerful tool. There are some small things in HCE that would have been nice if I could have done, including the ability to select more than one item from either the dendogram or the
graphs which were not directly next to each other (e.g. a selection while holding the control key in another application). It also would have been nice if you could cycle through the histograms a little easier. I would have loved the ability to be able to use both the up and down arrows as well as the right and left ones at the same time. Overall I was quite impressed with how HCE proved to be such an easy tool yet contributed greatly to the discovery process.
ERROR: undefined
OFFENDING COMMAND:

STACK: