Project #1: Human Height Data

In this project you will do a simple analysis of three datasets containing height data that you will find in the datasets folder on the Canvas site. The first dataset (hopeheights.dat) is heights of college students at Hope College. The first column of this data is the gender of the subject (1 female, 2 male) and the second column is the height in inches. See pythonstart.py to see how to load a file with multiple columns. Alternatively, you can tell numpy.loadtxt which columns of the datafile you want to load. Google numpy.loadtxt to see how to do this. The second data set (BBhtwt.dat) contains heights (in inches) and weights (in lbs) of professional baseball players. The third data set (Bigheightwt.dat) is the heights (in inches) and weights (in lbs) of 25,000 adolescents and various ages and both genders. This assignment is to first calculate the mean and the standard deviations for the height data in each of the three data sets. You are allowed to use standard functions such as np.mean() and np.std() to do this. For the Hope College data, also calculate a mean and a standard deviation for each gender separately. You can use a for loop with an if statement to create a separate list for each gender. The append command will be useful here as well. Report your results in a data table with a caption. Next, make a normalized histogram for the height measurements in each data set and try to fit a Gaussian function to the histogram. For the Hope College data make three histograms, one for the entire data set, one for males only, and one for females only. Your histograms should be included as figures with captions. In your discussion, first compare the means and the standard deviations that you calculated and discuss why they might have the relative sizes that they have, e.g. does it make sense that the means of the data sets increase in the order that they do? Finally, discuss the distributions of the data sets seen in the histograms. Which circumstances lead to height being distributed in a Gaussian distribution? How do these circumstances apply to each data set?

Project reports should be written in TeX and figures should be embedded in the text. Reports should have an Introduction giving background on the datasets being analyzed, an Analysis section talking about what methods were used to analyze the data and giving results, and a Discussion section where you discuss your results.