Title

Big data part 2

Problem Description

Download the file “bigData.txt”. It will be our input file for this problem.

Warning: bigData.txt is too large to be opened by Notepad.

Recall from the previous problem that worked with this data that the data is organized in two columns separated by a space. You do not know how many rows the data has (hint: it’s a lot!).

The first three lines of the file are shown below:

```
0.81472 5.4498

0.90579 14.083

0.12699 4.4615
```

Restrictions: You cannot use arrays for this question

Note: we are going to be using an equation you may not be familiar with. When writing code you may be asked to implement an equation or an algorithm that you don’t fully understand. This is OK. The hard part is going from the mathematical description (the equation) to code. You don’t need to understand where the equation came from, although it certainly helps.

Note 2: this problem is hard ☺

A common measure of the shape of a set of data is the standard deviation, σ. Briefly, the standard deviation is a measurement of how likely the data is to deviate from the mean. See https://en.wikipedia.org/wiki/Standard_deviation for a good description with pictures. For our data, it can be computed with the following formula:

$$\sigma = \sqrt{\left(\frac{1}{N}\sum_{\{i=1\}}^{N} x_i^2\right) - \left(\frac{1}{N}\sum_{\{i=1\}}^{N} x_i\right)^2}$$

Where x_i is the i^{th} point of data. For example, if we are figuring out the standard deviation for the 2nd column and are reading in the 3rd row, then $x_3 = 4.4615$ for the data we’ve generated.

Determine the following for each column of data:

1) The standard deviation
2) The maximum value in that column
3) The minimum value in that column

Then, determine the minimum and maximum of the whole data set.

Output all results to the screen

Testing

$$\sigma_{column\ 1} \approx 0.29$$

$$\sigma_{column\ 2} \approx 10$$

$$\min_{column\ 1} \approx 5 \times 10^{-7}$$

$$\max_{column\ 1} = 1$$

$$\min_{column\ 2} \approx -52.2$$

$$\max_{column\ 2} \approx 46.5$$

Time Target

***	less than 30 minutes
**	30-60 minutes
*	greater than 60 minutes

Section

File IO