Assignment #2

Instructions:

1. ALL DUE TIMES ARE IN EST
2. Upload Python files to this BB Assignment**.**
3. You are limited to 2 submissions
4. If you cannot submit the assignment by the due date, you can still submit it by the two days with a 25%penalty.
5. Any submissions after two days of the assignment class will be given a zero.

You must use a list to store the data

- In this assignment you are asked to calculate descriptive statistics for numbers found in columns of a given input file. The input files are tab-delimited files, which have six columns, you are to open the file and gather descriptive statistics on the column specified. The data files for this program can be found here, Right click, and "Save Link As.." on a Mac, or go to the page and save as the text:
- Data_file
- Data_file2
- Data_file3

  Modules you can use:

  sys

  math

  You cannot use any other Python modules, doing so will result in a zero for the assignment.

Name your program: **stats_in_python.py**

Notes about input file

- Some fields might not be numeric, you have to ignore these cases. If specified column contains only non-numeric then the program should raise appropriate message.
- Some of the numbers are NaN (Not a Number). When calculating descriptive statistics you have to ignore these numbers.
- You have to test your code with multiple examples. You can generate the output below from the data files provided.

pseudo code for your program

1. open the file
2. loop over data file and store the data in a list called **numbers**.
3. To loop over a file, you can use something like the following.

```
with open (file, 'r') as infile:
  for line in infile:
    # process input here
```

4. To split the line you're looping into see something like: num = **line**.split("\t")[2] # change the index 2 to something different
5. The column number should start with 0
6. Only store numeric values, and store them as floats. Use try .. except if this conversion does not happen correctly, see ValueError
7. When using split, make sure to use try .. except if an index goes out of bounds (see IndexError) and **exit** the program
8. Calculate descriptive statistics for the numeric floats in the list, and store into meaningful variable names
9. Print out the data (see examples below)

Use the following formulas for your descriptive statistics (n is the number of measurements):

Average: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Standard deviation: $$s = \sqrt{S^2}$$

Median: Sort the numbers first and find the middle element if number of elements is even, the median is the average of two middle number.

For example: Values for can be calculated using the equations, using the 65, 69, 70, 63, 70, 68 as example data:

Number of measurements: $n = 6$

Average: $\overline{x} = \frac{65 + 69 + 70 + \ldots}{6} = 67.5$

Variance: $s^2 = \frac{1}{6-1} \cdot \left((65 - 67.5)^2 + (69 - 67.5)^2 + (70 - 67.5)^2 + \ldots\right) = 8.3$

Standard deviation: $s = \sqrt{8.3} = 2.88$

Median

65, 69, 70, 63, 70, 68 -> 63,65,68,69,70,70

median is: (68 + 69) /2 = 68.5

Results format:

data_file.txt

Column: 3

Count = 125.000
ValidNum = 125.000
Average = 57.440
Maximum = 97.000
Minimum = 16.000
Variance = 308.490
Std Dev = 17.564
Median = 58.000

---

data_file2.txt

Column: 3

Count = 125.000
ValidNum = 124.000
Average = 57.621
Maximum = 97.000

Minimum = 16.000
Variance = 306.871
Std Dev = 17.518
Median = 58.000

---

data_file3.txt

Column: 3

Error: There were no valid number(s) in column 3 in file: data_file3.txt

---

data_file3.txt

Column: 10

Exiting: There is no valid 'list index' in column 10 in line 1 in file: data_file3.txt

---

data_file3.txt

Column: 1

Count = 2.000
ValidNum = 1.000
Average = 8.000
Maximum = 8.000
Minimum = 8.000
Variance = 0.000
Std Dev = 0.000
Median = 8.000

---

data_file2.txt

Column: 4

Count = 125.000
ValidNum = 124.000
Average = 0.821
Maximum = 0.940
Minimum = 0.640
Variance = 0.006
Std Dev = 0.078
Median = 0.840