**CSC 630 Introduction to Data Science**

**Assignment: Calculate Statistics**

**Background**

Exploratory data analysis (EDA) requires that we become familiar with the statistics describing our data. It’s also necessary to be able to read files and extract data to analyse.

**Procedure**

Sonnets.txt is a file holding the text of all of Shakespeare’s sonnets. For full credit, write a program to read Sonnets.txt and calculate the mean, median, and standard deviation of the number of words per sonnet. Alternatively, for a 10% penalty, you can read Sonnet\_lengths.csv, which holds the number of words in each sonnet. If you read Sonnets.txt, you’ll have to parse the text from the file to separate the sonnets. You can use read\_sonnets.py as a start for your program. If you read Sonnet\_lengths.csv, use the csv module to read the file. Whichever file you use, your program should print the mean number of words per sonnet, the standard deviation, and the median number of words per sonnet. You can use the formulae given in chapter 3 of the text – but I suggest you learn about the Python statistics module and use its mean(), median(), and stdev() functions. Round mean and standard deviation to one decimal place.

Your program should be general, in the sense that it should work for any file in the same basic format as Sonnets.txt. For example, it should work correctly if I run it over a file holding the short stories of Edgar Allen Poe, with each story separated by an integer by itself on a line. I’ll test your program by running it over files with similar formats. If you use Sonnet\_lengths.txt, I’ll try different, but equivalently formatted, csv files.

**Deliverables**

Name your program Asn1\_username.py, where username is your MSU user name and turn it in on Blackboard. For example, I would name my program Asn1\_ls555.py. **Do your own work – your program will be checked for originality by software designed to compare programs.**