Week 3 Assignment: Python

1) Install Anaconda/Python3 and the Spyder IDE

2) Install packages numpy, pandas, matplotlib.pyplot, scikit-learn, and any other packages that you wish to use.

3) The goal of this assignment is to practice writing Python code and to practice with reading in data, cleaning data, using pandas and numpy, and doing basic analysis.

Requirements:

1) Use the dataset called **StudentData2.csv**

2) Read this data into Python using pandas as a dataframe.

3) Clean the data according to these rules:

- The ID must be 6 numbers in length (no more and no less). Remove rows with IDs that are not 6 in length.

- The gender is 1 for male and 2 for female. No other gender values are permitted. Remove rows with genders other than 1 or 2.

- The age must be between 18 and 80. Remove rows with ages outside of this range.

- The gpa must be between 0 and 4.0. Remove rows with incorrect or missing gpa values.

- The sat score must be between 0 and 1600. Remove rows with sat missing or out of range.

- There are 5 class sections (1 – 5). Remove rows with sections that are missing.

- The final can be between 0 and 100. Remove any incorrect or missing values (remove the row).

- The project can be between 0 and 100. Remove any incorrect or missing values (remove the row).

4) Once the dataset is clean, print it **and** write it to a file called OUTFILE.txt.

5) Using numpy - and/or python statistical methods - perform the following statistical tests on the data:

(a) Is there a significant difference in final scores between the two gender groups? (Run an independent samples t-test and explain the p value and result. **Write** all results and explanations to the OUTFILE.txt.

(b) Is there a correlation between the final and the project? What is the r value? Write the results and explanations to OUTFILE.txt.

(c) Using ANOVA, is there a significant difference between the 5 sections for final score? Create box plots to visualize the 5 sections for final score. RE: http://www.marsja.se/four-ways-to-conduct-one-way-anovas-using-python/

Be sure that your boxplots are also copy and pasted to the Word doc that contains all the plots for this assignment.

(d) Using numpy, get the mean, median, max, min, var, and std dev for the final, the project, the gpa, and the sat. Write all the results to the OUTFILE.txt.

RE: http://www.scipy-lectures.org/packages/statistics/index.html

6) Using matplotlib, create two graphs: a scatterplot to look at the relationship between the final and the project scores, and a bar graph that shows the gpa values. Remember that to make a bar graph, you will have to categorize your data – do this in Python. Categorize (bin) the gpa into 5 groups: F, D, C, B, and A. Be sure that all graphs have titles and labels for x and y axes.

Deliverables:

1) For this assignment, you will create Python code that reads in a dataset and completes all the noted requirements. Create a zip folder. Into the folder, place the Python code.

2) Your program will generate (print and write) output and graphs. Write all printed output to a file called OUTPUT.txt and include this file in the zip.

3) Finally, your program will generate plots/graphs. Copy and paste all the plots that you create into a Word doc and include this Word doc into the zip.

4) Submit the zip via email to the TA and cc me. The title of the email MUST be Week 3 Assignment. Your name must appear in the email. All submitted code must run.

Remember – the goal of these programs is to help you to get used to using Python, pandas, numpy, and matplotlib.