Homework 1: Introduction to Python, Pandas, and Matplotlib

This assignment was adapted from one created by Dr. Steve Bogaerts.

Purpose

The purpose of this assignment is to practice using python and to begin to understand its data structures. This assignment will also introduce reading table data with the *Pandas* Python module and doing a simple plot with the *matplotlib* Python module.

Background and Preparation

Before attempting this assignment, please check your understanding and proficiency in Python. There are several posted resources available to you. These include some slides giving a summary of the Python language and a complete introductory text for Python.

You may also wish to review this comparison between Java and Python:
http://anh.cs.luc.edu/363/notes/JavaVsPython.html

Make sure you have *Pandas* (https://pandas.pydata.org/) installed. From a terminal, use your interpreter to install Pandas using the pip module.

On Windows: py -m pip install --user pandas

On Mac: python3 -m pip install --user pandas

Next, make sure to install *matplotlib* (https://matplotlib.org/). Similarly …

On Windows: py -m pip install --user matplotlib

On Mac: python3 -m pip install --user matplotlib

Task

Complete the following tasks.

1. Create a project folder for your homework project.
    a. This should contain your source code solution and your data files (described below).
2. Retrieve the adult census data from UC Irvine: http://archive.ics.uci.edu/dataset/2/adult
    a. Click on the download link and save the zipped folder.
    b. You will need the adult.data and the adult.names files. Place these files into your project folder under a folder called ‘data’. This is so you can access the files in your python strip by referencing the ‘data/adult.data’ file.
3. Download the starter code into your project folder.
4. The starter code contains a function called **getAgeList**.
    a. Use a search engine to find information on the read_csv function provided by Pandas.

   b. **WRITE** a docsctring in the solution file that explains what this function does in your own words. You must give a description of each of the named arguments in the read_csv function. This must be in your own words. Do not copy verbatim from the documentation page.
   c. Also in the docstring, **explain** the form origData.loc[:, 'age']. What does this line do?
5. **Create another function** called calcMean. This function should take a python list of numbers as input and return (do not print) the mean of the numbers (also known as the average).
   a. For example, calcMean([1,3,4]) should give 2.66666.
   b. Once your function is finished, modify main in the starter code to calculate and print the mean of the age data.
   c. When you use: print('The mean is: ', calcAverage(ages)), you should get approximately 38.581646755
6. Finally, have your script use matplot lib to **plot a histogram** of the ages values.
   a. You will use matplotlib's pyplot object to perform the plotting.
   b. Use a search engine to find a way to plot a histogram of the data. Your last line in main should be plt.show() to show the plot on screen.

Submission

Please zip your project folder and submit it to Moodle. Please name your zipped directory NAME_hw01.zip Where NAME is your name.

Evaluation

You will be evaluated using the following criteria:

- Organization
  - 2 points
  - Project folder is correctly organized and your script runs without errors
- Docstring
  - 3 points
  - Your docstring for getAgeList give a thoughtful description for the function and explains origData.loc[:, 'age'].
- calcMean
  - 5 points
  - Your function correctly calculates the mean of a list of values and returns a value.
- Histogram
  - 5 points
  - Your script uses matplotlib to visualize the distribution of ages using a histogram.
- Total points: 15