CIS 41A - Assignment 2

Write a program that analyzes scores of a sample group of students in online and hybrid sections of a class, and make recommendations for future students on whether their chance of success is better in the online or hybrid section.

Input files

The student score samples are in input2.zip as 3 files: data1.xlsx, data2.csv, and data3.csv. Your program should be prepared to read in one or more Excel or CSV files (minimum one file, maximum 20-30 files).

Each file is guaranteed to be in the following format:

* At least one row of data
* All rows of data in the file are valid, each row represents one student's scores in the class
* A row of valid data has 35 fields:
  + The first field is the letter 'O' for online or 'H' for hybrid
  + The rest of the fields are floating point numbers representing the scores that a student earn in the class
  + There is no empty field, all fields contain a floating point value except for the first field

The program

The program is already started in lab2.py. The file has a completely written main function and stubs of all the functions that you need to write. The main function calls 2 other functions:

1. readFile():

* Accepts a list of Excel or CSV files in the format described above.  
  When I test your program, I can enter one or multiple filenames, readFile should be able to handle a variable number of filenames.
* Returns 2 numpy arrays:
  + an array with all the scores of all the online students across all the input files
  + an array with all the scores of all the students in hybrid sections across all the input files
* The readFile function should call 2 other functions:
  + a readCSV() function that accepts a filename and returns a tuple of 2 list of lists, one for hybrid and one for online scores
  + a readExcel() function that accepts a filename and returns a tuple of 2 list of lists, one for hybrid and one for online scores

1. analyze():

* Accepts the two numpy arrays described in part 1
* Produces at least the 3 following data analyses:
  + A print out of the success rate of students in the online vs. hybrid data sets.  
    For each data set,  
     success rate = (num of students with a percentage grade of 70 or above) / total num of students in the set  
    The percentage grade for a student is the sum of all the scores for that student out of 519 possible points. For example, if the total score for a student is 482, then the percentage grade is (482/519)\*100 = 92.87
  + A histogram that clearly shows the distribution of percentage grades of hybrid vs. online data sets.
  + A line plot that clearly shows the mean of the hybrid vs. online assignment scores over time.  
    The first 8 fields of each line are the 8 assignment scores for one student. Find the mean of each of the 8 columns of scores for the online and hybrid data sets, then plot the mean values vs. time.   
    The assignment scores can be a way to track a student progress in the class during a quarter.

The recommendation

In addition to the 3 requirements above, feel free to add any other analyses to help you make the recommendation to students regarding choosing hybrid or online instruction. For example, the mean of the assignment scores is affected by the students who've effectively dropped the class by not turning in any more work, so maybe the standard deviation of the assignment scores should be looked at.

In a document, write a *short* paragraph for your recommendation. Is there a clear cut better choice for students or is it more of a "gray area"? Back up your assertion with specific observations from your calculated success rates, the histogram, line plot and any other data you may observe. Include a screen shot of your plots in the documentation, then convert the documentation to a pdf format.

Submit work

Turn in your lab2.py fle and your pdf file.

*[A footnote:*

*As you many know, data science or data analytics is a popular and booming field as businesses, industries, governments and education all move toward data driven business models. Data scientists are in demand because they have the capability to sift through massive amounts of data and observe patterns and trends and predict outcomes.*

*If you find that you enjoy sifting through rows and columns of scores to look for a possible trend while doing this assignment, perhaps data science is in your future. And Python will be right there with you.]*