Columbia Business School

Professor Kriste Krstovski
B9334 – Module 1

Spring 2019
Big Data for Finance

Homework Assignment #1

Due: 8:59 am, February 8th (Friday), 2019

- Upload your solutions to the Canvas system before the deadline (email Pu phe19@gsb.columbia.edu or Daheng dyang22@gsb.columbia.edu if that does not work for you). If you are uploading your solutions in separate files please package them into a single zip file and name it: *b9334_hw1_sol_first_last_name.zip* where you replace "*first_last_name*" with your full name. Also please make sure that your name is included as a comment at the beginning of each "*.py*" or "*.ipynb*" file that you submit. Lastly please make sure that you've also submitted a .pdf copy of those files.

- Homework assignments are to be done individually, and without the use of anyone else's solutions. You may obtain tips/tutorials from the internet, but soliciting help from others online or in person is not permitted. Cheating will strictly not be tolerated.

Question 1 (4 points)
In the lab session we covered two approaches for extracting variables of interest from a dataset stored in a csv file. The dataset at hand was the CRSP Quarterly Update Mutual Funds Dataset. The first approach was based on the "open" method which we use to read the file line by line. In our second approach we used the Pandas package.
Unlike the first approach which required more extensive code, when using Pandas we were able to complete the same task with only two lines of code. While more straightforward to implement we noticed that this approach required more time to execute.
To showcase the time difference between the two approaches we measured their execution time using the "time.clock" method. This method returns the current processor time. The execute time is computed by the taking the difference between the processor time after and before the execute of the code. The time comparison is provided in the "time_profiling.py" file. In this file we measured the execution time of the first approach when using list of lists to store the variables of interest. We noticed that this approach is almost 2 times faster than the Pandas based approach which proves the "no free lunch" theorem – while the approach using Pandas is more straightforward it does come with a price.
In this question you are asked to perform the same analysis using list of dictionaries rather than list of lists to store the variables of interest using the first approach. Comment on your findings and discuss the reason behind the time difference between the two data objects (list of lists vs. list of dictionaries). See if you can come up with a more time efficient code to extract these variables.

Question 2 (6 points)
During our lab session we went over an example task of extracting variables and values from a text file. As an example text file we used a schedule 13d filing from the SEC Edgar dataset. In this question you are

asked to write a Python code that would extract a set of variables (variables of interest) from the 2007 Microsoft Corporation annual report. You are given the original report which is in a Word document format (.doc file) and its plain text version (.txt file) which was converted from the original using Microsoft Word. The variables of interest are contained in the "FINANCIAL HIGHLIGHTS" table of this report. More specifically you are asked to extract the values of the following 3 variables from this table:
"Revenue", "Operating income" and "Total assets". The values in the table are given across 5 years (2007-2003 from left to right). You should use the original Word document version to get a better sense of the variables layout in this table and the plain text version to extract the variable values. For each variable use a dictionary to store its values across the years. Use the code example covered in the lab session as a starting point for developing your code solution. Your code should output the variable values.

Question 3 [Optional] (2 points)
In this question you are given the Compustat Annual Updates – Fundamentals Annual dataset in a csv file along with a pdf copy of its WRDS website that contains the variable descriptions. Both files are accessible through the dropbox folder that we used in our lab session:
https://www.dropbox.com/sh/f52fbtk51gbw4b3/AACSbtbs3WzYbra8huBzuMeFa?dl=0

Go over the list of variables and choose 20 variables that are of interest to you. Use the two approaches that we covered in our lab session to extract and store these variables and perform the same time analysis as in Question 1. More specifically, compute the execution time of first approach using list of lists and list of dictionaries, and the execution time when using Pandas. Comment on the observations.