Columbia Business School

Professor Kriste Krstovski
B9334 – Module 1

Spring 2019
Big Data for Finance

## Homework Assignment #2

**Due:** 8:59 am, February 15th (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 (6 points)

In class we covered various approaches for dealing with missing data. In this question you are given a modified version of "CRSP Mutual Funds - Monthly Returns and Net Asset Values" where some of the values are missing. The original dataset contains the following values:
`caldt` – Date (YYYYMMDD format)
`crsp_fund_no` - CRSP Fund Number
`mtna` - Total Net Asset Value
`mret` - Return per Share
`mnav` - Net Asset Value per Share

More about this dataset could found on the following WRDS webpage:
https://wrds-web.wharton.upenn.edu/wrds/ds/crsp/mfund_q/portnomap/index.cfm?navId=160

Using the approaches that we covered in class your task is to recover as much of the values as possible. The modified version of the dataset is stored in a file named "`monthly_return_mv.csv`". This file is accessible through the following dropbox:
https://www.dropbox.com/sh/f52fbtk51gbw4b3/AACSbtbs3WzYbra8huBzuMeFa?dl=0

The original version of the dataset contains values from the time period between 01.29.2010 and 09.27.2018. For each mutual fund, the data contains entries for each month of the year (105 entries). In addition, for each mutual fund the data is ordered based on the date. You may find this information useful when recovering missing values. Your code should output into a file the recovered entries for each mutual fund.

**Question 2 (4 points)**

In class we covered parallel processing along with high performance computing cluster. In addition, you were also given a tutorial on how to run job on the CBS computing cluster. Given that the "CRSP Mutual Funds - Monthly Returns and Net Asset Values" collection is fairly big (it contains 3,227,494 entries) you can take advantage of the fact that individual mutual funds could be processed independently and therefore speed up the task in Question 1 by splitting the file into smaller files and run your code on the compute cluster.

In this question you are asked to split the file into 10 files and run your code concurrently on the compute cluster using 10 jobs. Your code should print the mutual funds (CRSP Fund Number) that were processed. Each job on the cluster automatically generates a file that contains the standard output of that job. Submit the output files that were created by each job on the cluster along with the output files containing the recovered entries for each mutual fund