Data Programming Task for Causality Group

The next round of the interview process is an implementation task, where we look at the usage of good programming practices along with clarity, usability and stability of the resulted program code.

Please implement the below task in Python, using your favorite programming environment and any external libraries you find convenient. Do this on a platform which you can use later in our office with the same setup (your laptop or a cloud computer). Once satisfied and confident with the results, send us the final software code and your conclusions via email.

Implementing a data layer

Implement a data layer which downloads historical daily (open, high, low, close) prices and volume data from Yahoo Finance and Google Finance for given ticker symbols (e.g., AAPL). The data layer should download each data record exactly once from the Internet and store it locally for subsequent queries.

Calculating data fingerprints

Calculate the following fingerprints for Google's and Yahoo's data from 2016-01-01 to 2017-01-01 and for all ticker symbols of the S&P 400 Index:

1. The sample mean over all ticker symbols for each day and data field (prices and volume). Write the result into two files <source>__mean.csv (<source> is either google or yahoo) with columns date, open, high, low, close and volume.
2. The sample variance over all values between 2016-02-11 (including) and 2016-11-08 (excluding) for each ticker symbol and data field. Write the result into files <source>__variance.csv using the format of part 1.
3. Find the lowest and highest close price for each ticker symbol between 2016-01-18 (including) and 2016-10-18 (excluding). Write the result into files <source>__lowest_highest.csv with columns as ticker (ordered alphabetically), lowest_close, and highest_close.
4. The standard deviation for the difference of Google's and Yahoo's data between 2016-04-17 (including) and 2016-12-05 (excluding) calculated over all ticker symbol for each day and data field. Write the result into file google_yahoo__comparison.csv using the same format as for part 1.

Merging the data

Extend the data layer in order to provide consolidated data source, using both Google's and Yahoo's data. Consolidated data should contain the best quality data which is obtainable using these two sources. Available data, prices and volumes might be different in the two sources. When this happens you will have to chose between the sources, remove bad data points, take average of the two, etc… Any creative solution is welcome to arrive to the cleanest data with best coverage. Please describe your decisions and reasoning when choosing a filtering / selection method.

Then calculate fingerprints 1. - 3. for the consolidated data, and put the results into the files consolidated__mean.csv, consolidated__variance.csv, and consolidated__lowest_highest.csv, similarly as above. Finally, compare the consolidated data for both Google's and Yahoo's data as described by fingerprint 4., and provide the result in two files named consolidated_google__comparison.csv, and consolidated_yahoo__comparison.csv.

Results

Please send the final result to hiring@causality-group.com, including:

- Description of your decisions for the consolidated data source in the email body. This should be in a format which you would use to communicate challenges and decisions in a clear and concise manner in a working environment to your colleagues.
- All the csv files mentioned above attached in a zip file data_programming_results.zip.
- The program code attached in a zip file data_programming_code.zip.

Good luck,
Causality Group