FIN-221: Machine Learning in Finance
HW 2
Due on October 06, 2025

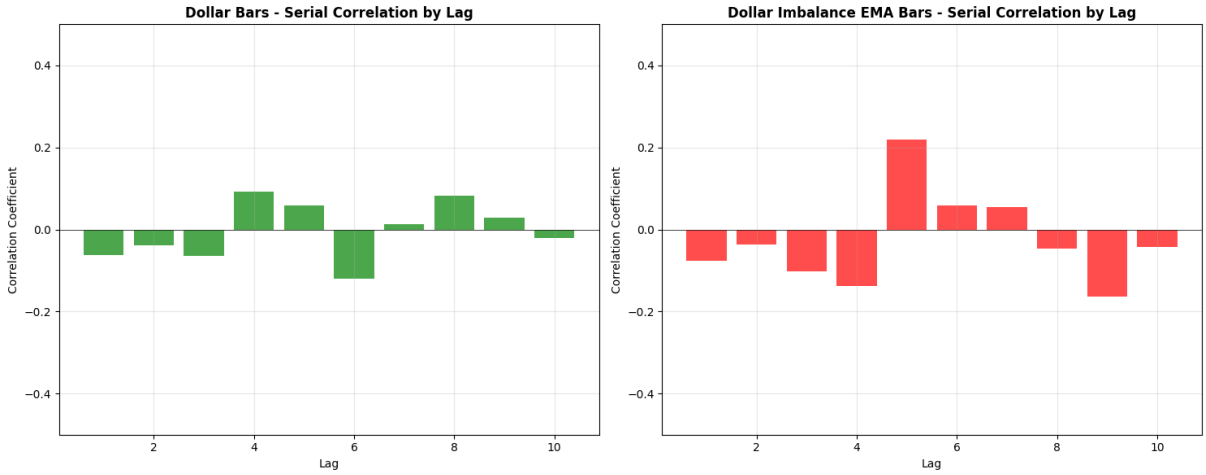
1. Exercise 2.2 of the textbook. You can use the programs and modules developed by Hudson and Thames. For futures roll data see https://raw.githubusercontent.com/hudsonand-thames/example-data/main/futures stitched.csv

   I have also uploaded a zip file of sample ES data (courtesy of Hudson and Thames).

Question 1 and 2 work are in GitHub -
https://github.com/McSavage/MLFinLab/blob/main/notebooks/FIN221_HW2_H%26A_1.ipynb

2.2 On a series of E-mini S&P 500 Futures tick data, compute dollar bars and dollar imbalance bars. What bar type exhibits greater serial correlation? Why?

At the first lag Dollar Bars show less serial correlation than Imbalance Bars; neither is big.



Serial correlation is rather subdued in both series.

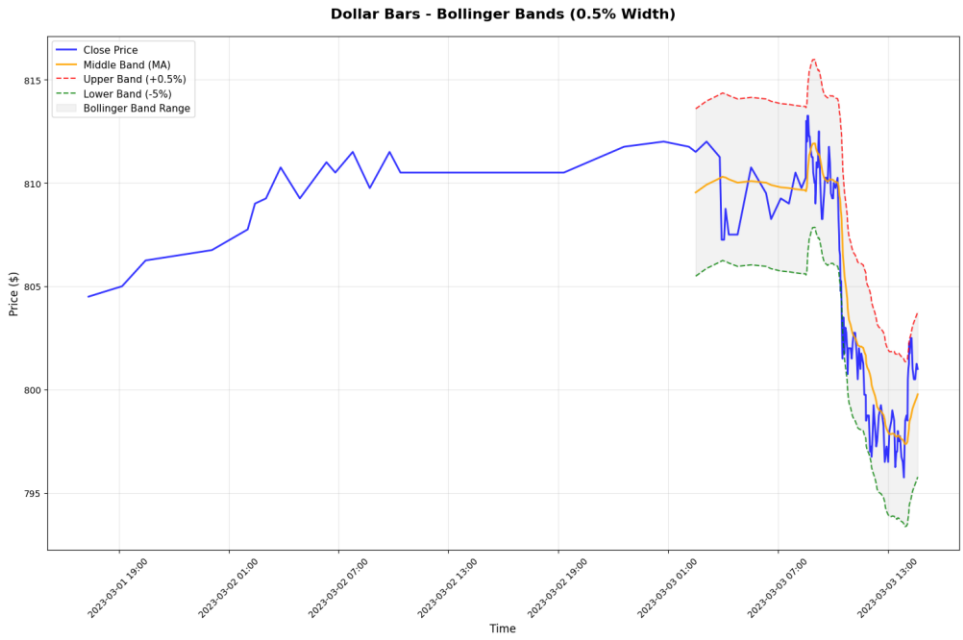
```
=== SUMMARY STATISTICS ===
Dollar Bars Returns:
  - Mean: -0.000026
  - Std: 0.001231
  - Skewness: -0.1557
  - Kurtosis: 1.0156
  - Number of observations: 163

Dollar Imbalance EMA Returns:
  - Mean: -0.000076
  - Std: 0.001248
  - Skewness: 0.0742
  - Kurtosis: 0.6968
  - Number of observations: 129

=== LJUNG-BOX TEST FOR SERIAL CORRELATION ===

Dollar Bars - Ljung-Box Test:
  Lags with significant serial correlation (p < 0.05):
    No significant serial correlation detected

Dollar Imbalance EMA - Ljung-Box Test:
  Lags with significant serial correlation (p < 0.05):
    No significant serial correlation detected
```

2. Exercise 2.4 of the textbook. Same goes here regarding the data.

2.4 Form E-mini S&P 500 futures dollar bars:

A. Compute Bollinger bands of width 5% around a rolling moving average. Count how many times prices cross the bands out (from within the bands to outside the bands).

   Here using a width of 0.5% - ...one half of one percent...

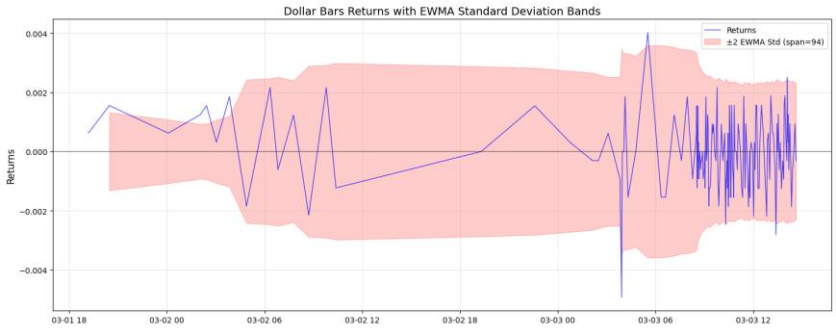


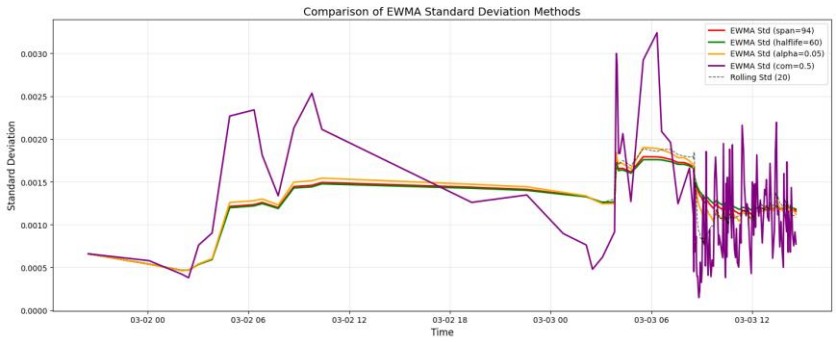

```
=== BAND BREACH ANALYSIS ===
Upper band breaches: 0 (0.0%)
Lower band breaches: 4 (2.8%)
Total breaches: 4 (2.8%)
```

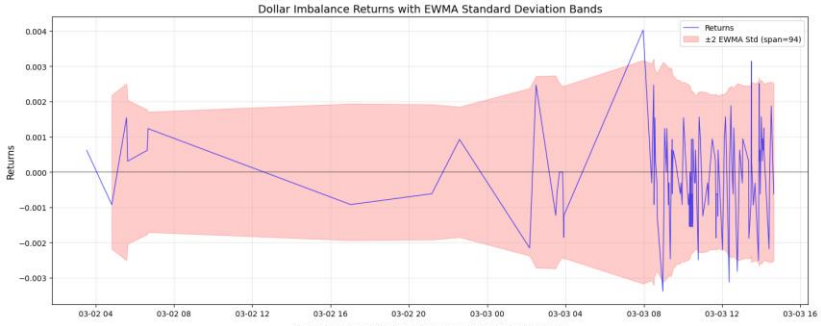
B. Now sample those bars using a CUSUM filter, where $\{y_t\}$ are returns and h = 0.05. How many samples do you get? - **9**

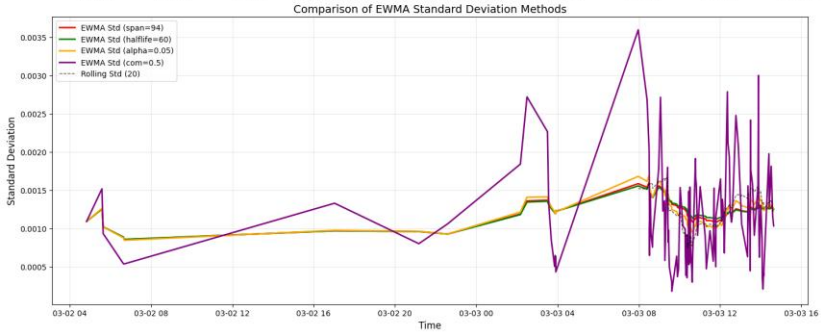
```
=== MLFinLab CUSUM FILTER ANALYSIS ===
Threshold (h): 0.005
Total T-Events detected: 9
Original dollar bars: 164
Sampling ratio: 5.5%
```

C. Compute the rolling standard deviation of the two-sampled series. Which one is least heteroscedastic? What is the reason for these results?
Imbalance bars look a little less heteroscedastic. Looks like it is suppressing some noise at the start of the day 03/03









3) Repeat Exercise 2.2 with Google data from https://github.com/jjakimoto/ nance ml/blob/master/datasets/Google.csv

Question 3 Work is in GitHub:

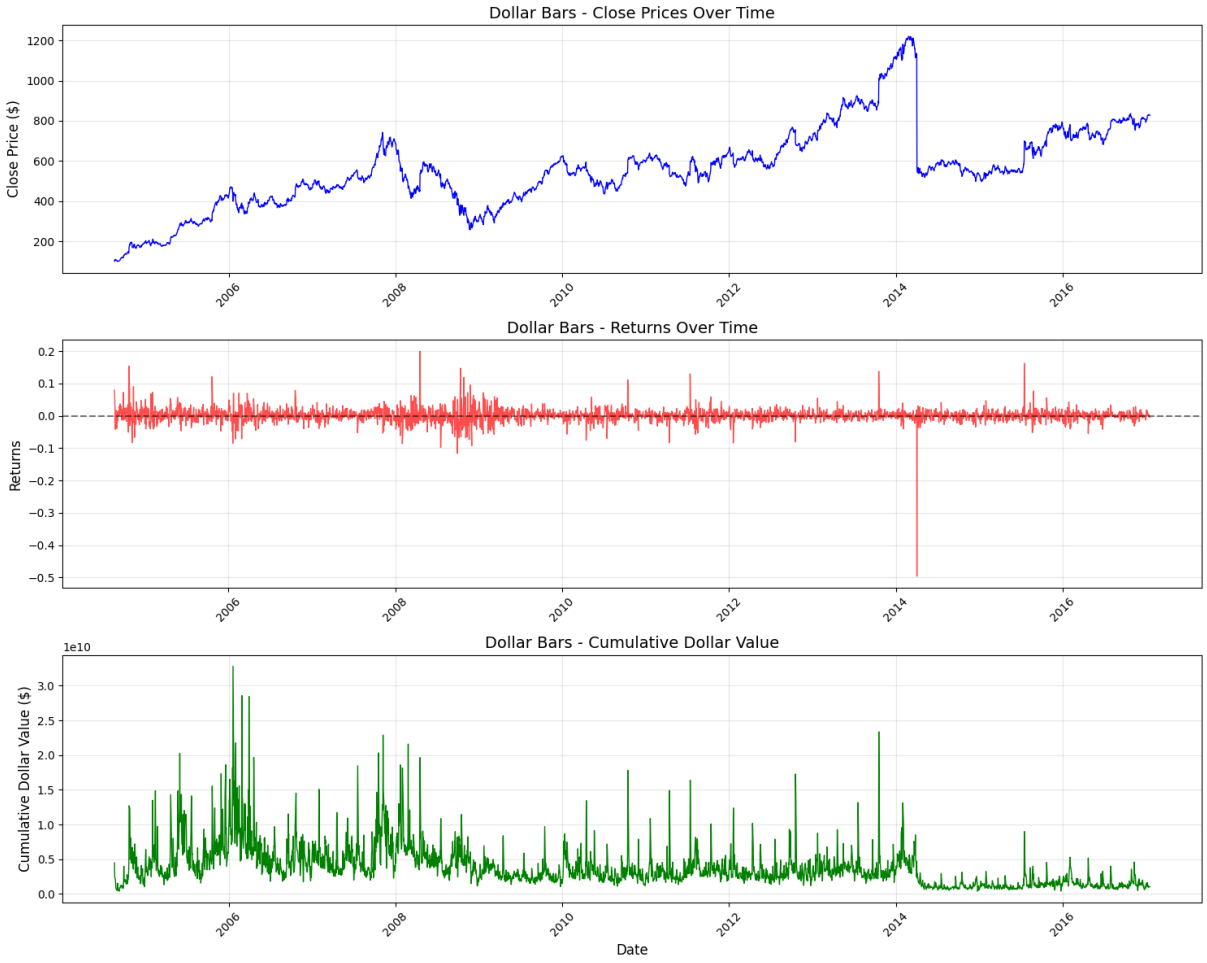
https://github.com/McSavage/MLFinLab/blob/main/notebooks/FIN221_HW2_Google.ipynb

This data set appears to be 13 years of daily bars for some security

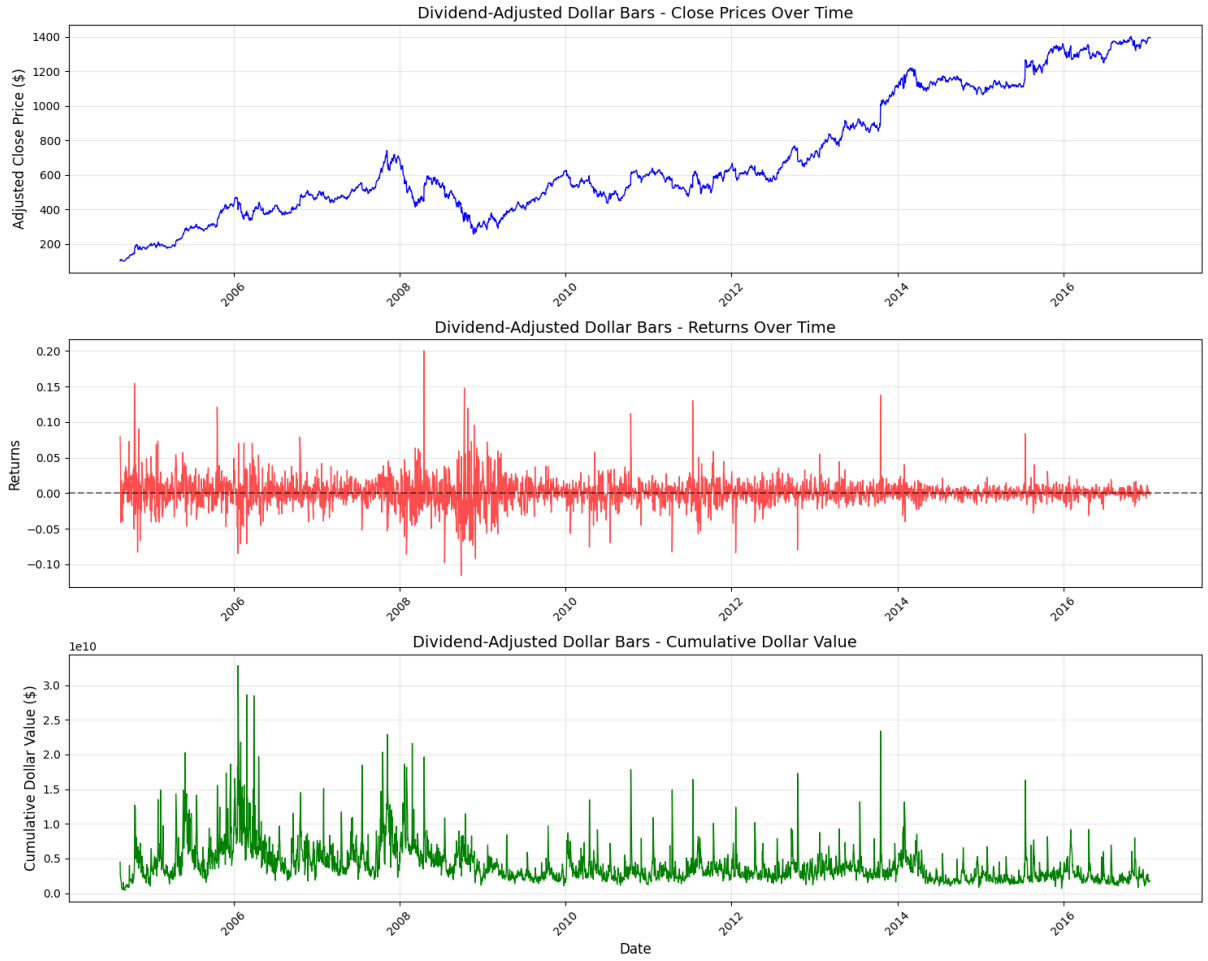
There is one dividend event and no splits

```
Ex-dividend dates and amounts:
  2014-04-03: $567.9717
```

That's the big move in the chart.



Adjusted series:



In these examples the original data as tick bars and modified dollar bars don't have a lot of difference vis a vis serial correlation. Using avg_daily_dollar_volume for threshold doesn't add much info.

```
dollar_bars = standard_data_structures.get_dollar_bars(
    tick_df,
    threshold=dollar_threshold
)
```