FIN-221: Machine Learning in Finance

HW1

Due on September 22, 2025

Daniel Savage

1. Read up on futures from any resource you like. I recommend the relevant chapter in Hulls book, Options, Futures, and Other Derivatives.

    I have the 5th edition of Hull, it replaced the 1st edition I had in my office sometime in the early 2000s. I traded Stock Index Futures and Options for Timber Hill in the 1980s on the CME, CBOT and CBOE. I was also an NASDAQ Market Maker in the early 1990s at First Boston.

    I built a Counter Party Credit Risk Management System at Prudential Global Derivative in the early 2000s. Prudential cleared just about every listed commodity product globally.

    Most recently I worked on the US$ Swaps Pre-trade Pricing system for BofA including the transition from Libor (ED) to Sofr (SR3). Trade pricing for US$ Swaps, SOFR Futures, Bond Futures, and US Treasuries.

2. Exercise 2.1 of the textbook. You can use the programs and modules developed by Hud son and Thames. For futures roll data see https://raw.githubusercontent.com/hudson and-thames/example-data/main/futures stitched.csv I have also uploaded a zip le of sample ES data (courtesy of Hudson and Thames).

Work is in my GitHub:

https://github.com/McSavage/MLFinLab

2.1 On a series of E-mini-S&P 500 futures tick data:

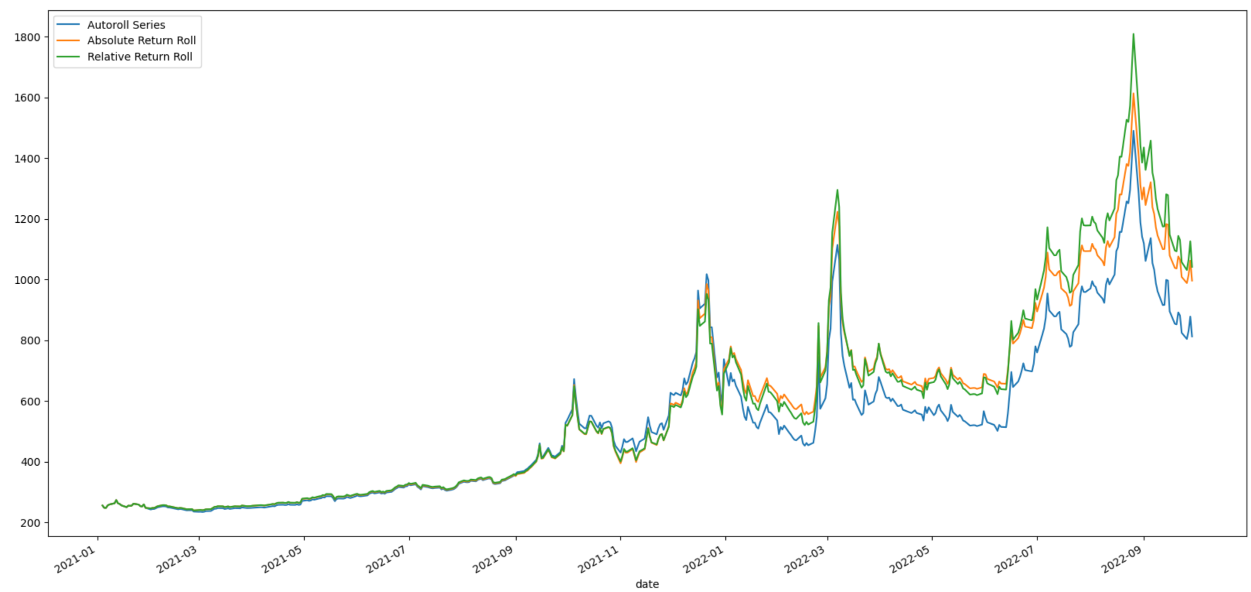
a. Form tick, volume, and dollar bars. Use the ETF trick to deal with the roll.

This data doesn't look like S&P futures data to me. But I believe this is the data set intended for this exercise. The levels are wrong, and S&P futures roll on the third Friday, not month end.

In hudson_example2.ipynb

Uses: mlfinlab.multi_product.etf_trick -> get_futures_roll_series

futures_stitched.csv data set



In addition to the adjusted close, I added volume and value fields.

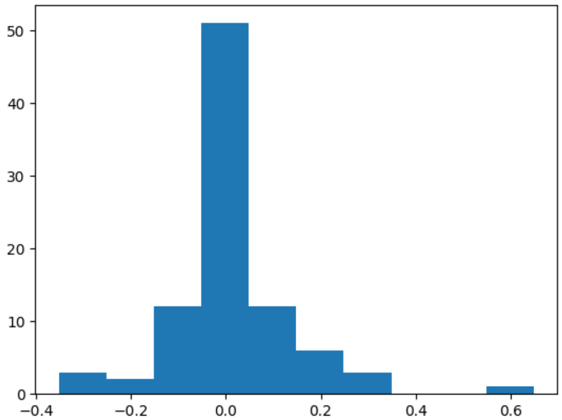
I dropped the 2 NA volume rows

I'll tread these daily observations as tick data.

I aggregated the daily observations to weekly "tick" bars:

continuous_contract_relative_method.groupby(pd.Grouper(freq=f'W'))

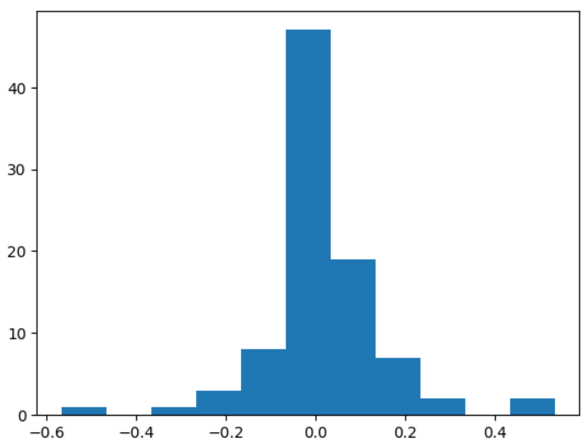
There are 91 weekly bars resulting. With return distribution:



I aggregated over volume using average weekly traded_volume = 13,722 contracts

continuous_contract_relative_method.groupby(bar(np.cumsum(continuous_contract_relative_method['volume']), traded_volume))

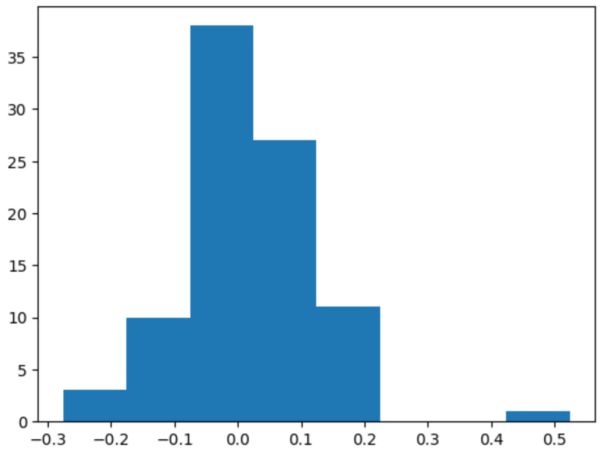
There are 91 weekly bars resulting. With return distribution:



I aggregated over volume using average weekly dollar market_value = 7,287,227

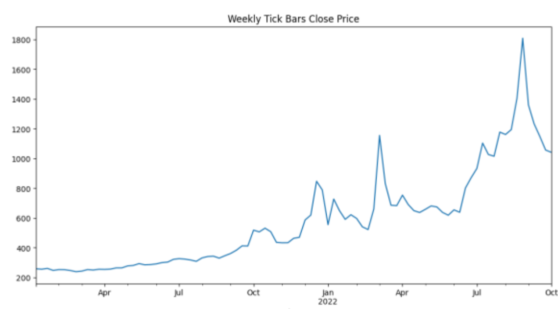
continuous_contract_relative_method.groupby(bar(np.cumsum(continuous_contract_relative_method['value']), market_value))

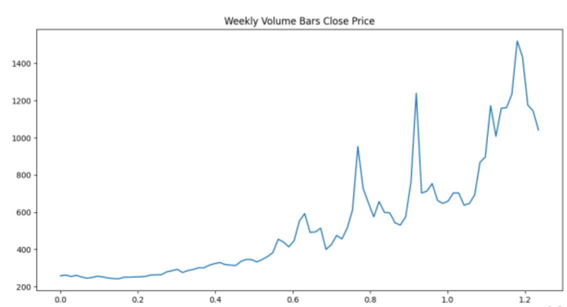
There are 91 weekly bars resulting. With return distribution:

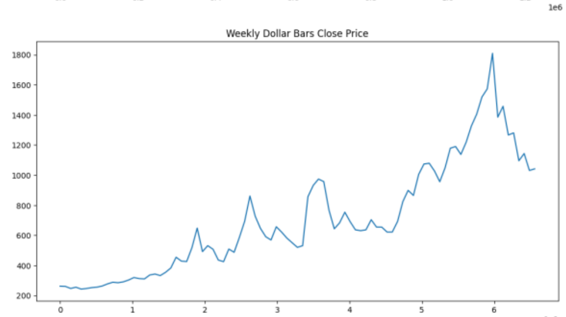


b. Count the number of bars produced by tick, volume, and dollar bars on a weekly basis. Plot a time series of that bar count. What bar type produces the most stable weekly count? Why?

The construction resulted in each bar type produced 91 observations

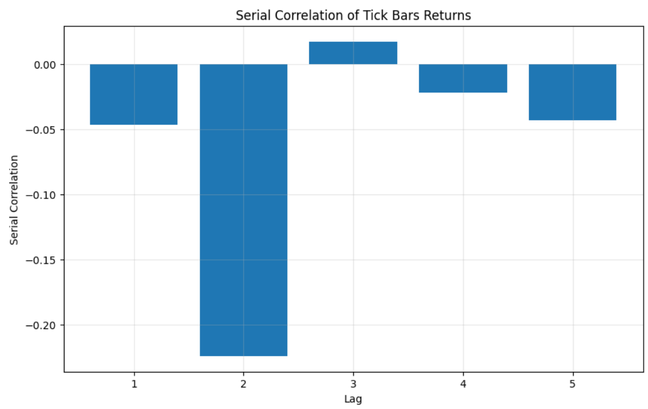






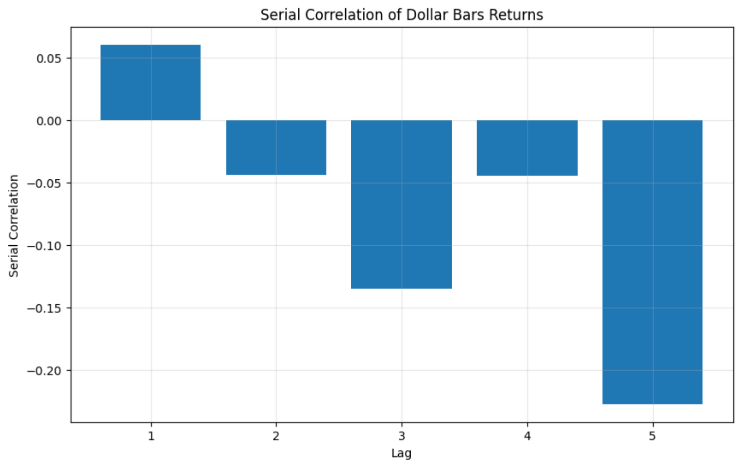
c. Compute the serial correlation of returns for the three bar types. What bar method has the lowest serial correlation?

Serial Correlation (Lag-1) of Tick Bars Returns: -0.0467 <- lowest
Serial Correlation (Lag-1): -0.0467
Serial Correlation (Lag-2): -0.2243
Serial Correlation (Lag-3): 0.0172
Serial Correlation (Lag-4): -0.0219
Serial Correlation (Lag-5): -0.0433



Serial Correlation (Lag-1) of Volume Bars Returns: -0.1098
Serial Correlation (Lag-1): -0.1098
Serial Correlation (Lag-2): -0.1383
Serial Correlation (Lag-3): -0.0842
Serial Correlation (Lag-4): 0.0096
Serial Correlation (Lag-5): -0.0555

Serial Correlation (Lag-1) of Dollar Bars Returns: 0.0607
Serial Correlation (Lag-1): 0.0607
Serial Correlation (Lag-2): -0.0436
Serial Correlation (Lag-3): -0.1346
Serial Correlation (Lag-4): -0.0446
Serial Correlation (Lag-5): -0.2270



d. Partition the bar series into monthly subsets. Compute the variance of returns for every subset of every bar type. Compute the variance of those variances. What method exhibits the smallest variance of variances?

Sorry, I’m tired.

e. Apply the Jarque-Bera normality test on returns from the three bar types. What method achieves the lowest test statistic?

Jarque-Bera Normality Test Results:
==

Tick Bars:
  JB Statistic: 91.8105
  P-value: 0.0000
  Normal: No
  Skewness: 0.7411
  Kurtosis: 4.7208

Volume Bars:
  JB Statistic: 122.5846
  P-value: 0.0000
  Normal: No
  Skewness: -0.1052
  Kurtosis: 5.7136

Dollar Bars:
  JB Statistic: 27.6523 <- lowest
  P-value: 0.0000
  Normal: No
  Skewness: 0.4402
  Kurtosis: 2.5688

3. Read the paper The Volume Clock: Insights into the High Frequency Paradigm, David Easley, Marcos M. Lopez de Prado, Maureen OHara and answer the following questions:

- What is the volume-clock metric and why is it important?
    - Humans look at security price series in time buckets. Perhaps viewing returns from the last trade in a one-minute (hour, day, week, month, etc.,) interval to the last trade in the next one-minute (hour, day, week, month, etc.,) interval. Trade count, trade volume, VWAP etc., do not enter consideration or are of secondary interest. Traders may look at candle-stick charts (open, high, low, close) to extract trends with moving averages still looking at close to close prices.
    - The problem with close-to-close price returns is that they are very much non-normal – exhibiting leptokurtosis (fat tails) and non-stationarity in mean and variance
    - The volume-clock metric moves the analysis from time-based metrics to event-based metrics. Specifically, the volume-clock measures time in terms of trading volume or transaction counts.
    - Clark [1973] expands on Mandelbrot [1967] idea as a solution to “recover normality” in return data.
    - The ideas (volume clock) gain traction with the advent of HFT in the 1990s
- What is the main thesis of the paper regarding HFT vs LFT?

    - HFTs are characterized in the paper as predators using their informational advantage in market microstructure to identify LFT and exploit their predictable behavior.
    - LFTs leak information about their action and the HFT react to this information leakage to anticipate the LFT's actions to exploit market microstructural opportunities.
- Explain how the authors have formed Figure 4 and how they interpret the results.
    - The authors start with a data set containing one year of E-mini futures trade between 11/7/2010 and 11/7/2011. I believe that they have quantity and date/time to the second of all trades. They cross tabulate the data bucketing volume data for each second (0 to 59) of a minute (for all minutes in the hour) vs hour (0 to 23) in the day (ignoring the date of observation) and calculating for each bucket the percentage of second buckets' minute volume.