Coding Assignment #1

1. Load data from the SPY.csv file and NAESX.csv file.
2. Look around to see if there are any differences, if so, how?
3. Compute 1-day return based on the “Adj Close” column. Also compute the open-to-close (“day”) returns and the close-to-open (“night”) returns for both files. Note the night time return is computed differently due to price adjustments. (use .shift())
4. Create a new column named “EOM3” that is equal to 0 on all days except for the third business day before the end of the month. On those days, the EOM3 column should equal 1. For example, if the last three business days of the month are the 27th, 28th, and 29th, then the 27th would have EOM3=1 and all other days will have EOM3=0. (this will require a groupby, .nth()).
5. Compute the average return (close-to-close, day, and night) over all days and only over the EOM3=1 subset. Determine which strategy would be better, either buying and selling every day, or only buying and selling on the third business day before the end of the month.
6. Create a portfolio, which will be a new dataframe, where you hold some percentage of SPY and some percentage of NAESX, you decide. Then compute the MONTHLY returns of your strategy. (this will require merging; to compute monthly returns of daily data you can use groupby and .cumprod() or .prod() method)
7. Run a linear regression, using StatsModels or SkLearn to forecast next month’s return of your portfolio using a two day (you could do longer if you want) lag of previous returns. Be sure to create a train and a test set, you should also create a validation set to do cross-validation (i.e. backtest). (this will require you to create a feature matrix where the features are the lagged returns, and a response vector which is the returns). Note a typical k-fold cross validation will not work on time series because of the ordering of time!
8. Evaluate performance of this predictor on a validation or test set using the square root of the mean square error and R^2.