Pairs Trading Project: Large-Scale Cointegration Testing Across Five Equity Sectors

Chun Samuel, Padua Francis
Dr. Doina Bein

Table of Contents

Introduction

What is Pairs Trading?

Pairs trading is a market-neutral strategy designed to earn profit regardless of broad market direction by exploiting temporary divergences between two historically co-moving securities.

Origin

This strategy was developed at Morgan Stanley in the mid-1980s, by a group of quantitative analysts led by Gerry Bamberger and Nunzio Tartaglia. The pairs strategy was motivated by the mean-reversion theory—a financial hypothesis that theorizes that asset prices and historical returns eventually return to their long-term mean over time.

Inspired by this, they selected co-moving stocks—that were from the same sector (companies that share similar business activities or are in the same industry)—by relying on a simple correlation coefficient to gauge if the pair was worthy of employing the pair trading strategy (later, cointegration was used as a more thorough test for employing pairs trading).

How the Strategy Works

1. Identify two stocks whose prices move together (high correlation and, ideally, cointegration).
2. Monitor the spread between their log prices.
3. When the spread deviates significantly from its historical mean, short the “rich” leg and buy the “cheap” leg.
4. Close both positions when the spread reverts, locking in the convergence profit.

Overview

This is a student-led research project facilitated by a summer research opportunity under grant-funded project CIC | PCUBED (Center for Inclusive Computing | Pathways, Pipeline, Practice) at California State University, Fullerton (CSUF). This project explores pairs trading opportunities of 45 stocks in each of the five different chosen equity sectors (Technology, Pharmaceuticals, Energy, Insurance, Fast Food). The main goal of this project is to understand the processes that buttress the pairs trading strategy—screening for pairs worth testing, testing cointegration, and running a simple backtester to see the profitability of a selection of pairs using historical stock data.

The requirements of this research as specified from CIC | PCUBED are:

1. *Choose and download at least two stock price values from MarketWatch, Nasdaq or other Internet websites that provide free download.*
2. *Apply one or more of the Greedy algorithms mentioned in the lectures for computing the mean and the spread as the range, variance, the standard deviation, and/or IQR, on a pair of stock values downloaded from MarketWatch, Nasdaq or other Internet websites.*
3. *Compute the correlation, and if you have time, compute the cointegration, on that pair of stock values.*
4. *Code using C/C++ or Python to compute the above data.*
5. *Provide results and conclusion to this work.*

This project expanded the scope from finding the correlation of only one pair to finding correlation using C++ and finding cointegration using Python libraries of approximately ~1980 pairs (45 permute 2).

Objectives

The objectives with this research project are:

- Data Exploration*: Fetch historical price data (≈1 year, daily) for all stocks and compute summary statistics and Pearson correlation for every pair within each sector.*
- Statistical Testing*: Identify candidate pairs that are highly correlated and test each for cointegration using the Augmented Dickey-Fuller (ADF) test (Engle-Granger two-step method).*
- Strategy Backtesting*: For pairs deemed cointegrated, perform a simple mean-reversion trading backtest. The strategy uses the log price residual between two stocks, enters positions when the residual's z-score hits extreme thresholds, and exits when it reverts toward the mean.*
- Performance Evaluation*: Assess annual returns, Sharpe ratio, number of trades, and other metrics for each pair's strategy, and analyze whether any consistently profitable, market-neutral opportunities exist.*
- Learnings and Next Steps*: Reflect on findings, discuss limitations (e.g. transaction costs, assumptions) and propose improvements (like dynamic hedging, more rigorous tests) for future research.*

Hypothesis

Equity pairs that pass a 5%‑level cointegration test (ADF) will, on average, generate positive risk-adjusted returns when traded with a simple mean-reversion strategy (±1.5 σ entry, ±0.5 σ exit) over the following year, whereas non-cointegrated pairs will not.

Methodology

Collecting Data

To collect our data, we utilize python library *yfinance* to scrape stock data into CSVs and store them in a directory, saving columns Date, Open, High, Low, and Close. To proceed with the program, we also guarantee that all date entries are common in all CSVs. We do so by performing a set intersection on all CSVs for 'Date' entries.

Correlation Analysis

Before we continue with correlation analysis, we use C++ to first guarantee that all date entries are the same in all CSVs. We do so by performing a set intersection on all CSVs for 'Date' entries to ensure integrity of data:

$$\begin{array}{lll} \text{Given} & D = \{D_1, D_2, D_3, ..., D_n\}, & \text{Family of sets of dates, one per CSV} \\ \text{Then} & S = \bigcap_{i=1}^{n} D_i, & \\ \text{if} & \forall i \; (|S| = |D_i|), & \text{then we continue with the program} \end{array}$$

Then, for every stock, the Pearson correlation coefficients are calculated as a simple screening to sift for the best candidate pairs. Without screening, in the practical world, calculations would be forced to run on every possible pair—which requires excessive computational power for little guarantee of profitability. A high Pearson correlation coefficient will signal, on a basic level, how closely stocks move together. Low correlation thus assumes that such a pair is less likely to be good trading candidates.

Pearson correlation is calculated not on the raw price, but instead on the log prices. This is because there is a fatal error in executing calculations on raw prices—one stock, which may have mostly large raw prices, may change by larger prices than the other stock's smaller price changes. Hence, log prices are used, as log prices make these differences linear—it grants the ability to measure by percentage movements rather than raw value movements.

Performing the Pearson correlation, we take the log prices of both the first stock's raw prices A and second stock's raw price B to perform:

$$r = \frac{\sum_{i=1}^{n}(\ln A - \overline{\ln A})(\ln B - \overline{\ln B})}{\sqrt{\sum_{i=1}^{n}(\ln A - \overline{\ln A})^2}\sqrt{\sum_{i=1}^{n}(\ln B - \overline{\ln B})^2}}$$

If $r > 0.7$, then the program concludes that this pair is correlated.

However, high correlation alone is an inadequate measure to see if a candidate pair is viable for pairs trading. Correlation captures short-term co-movement but not whether the spread between two price series will mean-revert. It is conceivable that a pair of stocks may trend together, yet drift further and further apart. Therefore, we proceed to test for cointegration for a more robust analysis of the relationship of said pair. [insert a picture of an example]

Cointegration Testing

To find pairs with a mean-reverting relationship, we employ the Engle-Granger method, a two-step method to test whether a pair is cointegrated, which 1) uses Ordinary Least Squares (OLS) to derive residuals, 2) which then runs the Augmented Dickey-Fuller (ADF) test on the residuals found previously to confirm stationarity.

From a First-Step Regression formula, we can derive a formula to extract residuals (also from log prices):

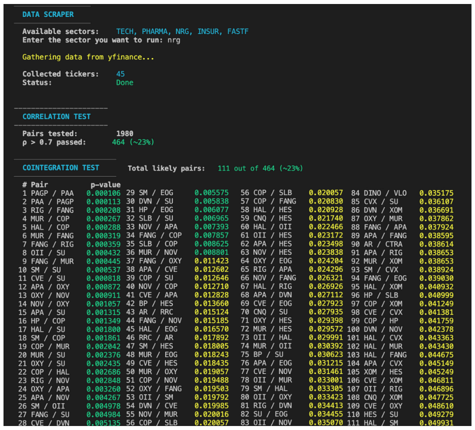
$$\ln B = \alpha + \beta \ln A + \epsilon_t$$
$$\epsilon_t = \ln B - \alpha - \beta \ln A$$

Using these residuals, we may then pass them through the ADF test. ADF confirms stationarity by using statistical hypothesis testing—the null hypothesis of ADF is that the series has a unit root (non-stationary), and the alternative hypothesis stating that our residuals are stationary, returning a p-value. This p-value returns the probability of the likelihood of the null hypothesis. Given this p-value falls below the chosen threshold of 5%, then there is at least 95% confidence that the null hypothesis is unlikely to be true, thus presenting evidence to support the alternative hypothesis. As a result, we can conclude this pair is stationary, which is therefore cointegrated. in which we find if 3) the null hypothesis is rejected, then the pair is cointegrated. If not, it is not cointegrated.

Results

Figures

Below are some figures from running our program.



DATA SCRAPER

Available sectors: TECH, PHARMA, NRG, INSUR, FASTF
Enter the sector you want to run: nrg

Gathering data from yfinance...

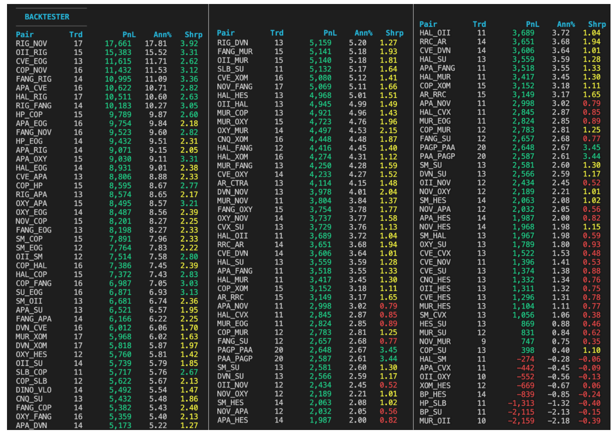
Collected tickers: 45
Status: Done

CORRELATION TEST

Pairs tested: 1980
ρ > 0.7 passed: 464 (~23%)

COINTEGRATION TEST Total likely pairs: 111 out of 464 (~23%)

# Pair	p-value						
1 PAGP / PAA	0.000106	29 SM / EOG	0.005575	56 COP / SLB	0.020057	84 DINO / VLO	0.035175
2 PAA / PAGP	0.000113	30 DVN / SU	0.005838	57 COP / FANG	0.020830	85 CVX / SU	0.036107
3 RIG / FANG	0.000208	31 HP / EOG	0.006077	58 HAL / HES	0.020928	86 DVN / XOM	0.036691
4 MUR / COP	0.000267	32 SLB / SU	0.006965	59 CNQ / HES	0.021740	87 OXY / MUR	0.037862
5 HAL / COP	0.000288	33 NOV / APA	0.007393	60 HAL / OII	0.022466	88 FANG / APA	0.037924
6 MUR / FANG	0.000319	34 FANG / COP	0.007857	61 OII / HES	0.023172	89 APA / FANG	0.038595
7 FANG / RIG	0.000359	35 SLB / COP	0.008625	62 APA / HES	0.023498	90 AR / CTRA	0.038614
8 OII / SU	0.000432	36 MUR / NOV	0.008801	63 NOV / HES	0.023838	91 APA / RIG	0.038653
9 FANG / MUR	0.000445	37 FANG / OXY	0.011423	64 OXY / EOG	0.024204	92 MUR / XOM	0.038653
10 SM / SU	0.000537	38 APA / CVE	0.012602	65 RIG / APA	0.024296	93 SM / CVX	0.038924
11 CVE / SU	0.000818	39 COP / SU	0.012646	66 NOV / FANG	0.026321	94 FANG / EOG	0.039030
12 APA / OXY	0.000872	40 NOV / COP	0.012710	67 HAL / RIG	0.026926	95 HAL / XOM	0.040932
13 OXY / NOV	0.000911	41 CVE / APA	0.012828	68 APA / DVN	0.027112	96 HP / SLB	0.040999
14 NOV / OXY	0.001057	42 BP / HES	0.013660	69 CVE / EOG	0.027923	97 COP / XOM	0.041249
15 APA / SU	0.001315	43 AR / RRC	0.015124	70 CNQ / SU	0.027935	98 CVE / CVX	0.041381
16 HP / COP	0.001349	44 FANG / NOV	0.015185	71 OXY / HES	0.029398	99 COP / HP	0.041759
17 HAL / SU	0.001800	45 HAL / EOG	0.016570	72 MUR / HES	0.029572	100 DVN / NOV	0.042378
18 SM / COP	0.001861	46 RRC / AR	0.017892	73 OII / HAL	0.029991	101 HAL / CVX	0.043363
19 COP / MUR	0.002042	47 SM / HES	0.018005	74 MUR / OII	0.030392	102 HAL / MUR	0.043430
20 MUR / SU	0.002376	48 MUR / EOG	0.018243	75 BP / SU	0.030623	103 HAL / FANG	0.044675
21 OXY / SU	0.002435	49 CVE / HES	0.018435	76 APA / EOG	0.031215	104 APA / CVX	0.045149
22 COP / HAL	0.002686	50 MUR / OXY	0.019057	77 CVE / NOV	0.031461	105 XOM / HES	0.045249
23 RIG / NOV	0.002848	51 COP / NOV	0.019488	78 OII / MUR	0.033001	106 CVE / XOM	0.046811
24 OXY / APA	0.003260	52 OXY / FANG	0.019503	79 SM / HAL	0.033305	107 OII / RIG	0.046896
25 APA / NOV	0.004267	53 OII / SM	0.019792	80 OII / OXY	0.033423	108 CNQ / XOM	0.047725
26 SM / OII	0.004978	54 DVN / CVE	0.019985	81 RIG / DVN	0.034413	109 CVE / OXY	0.048610
27 FANG / SU	0.004984	55 NOV / MUR	0.020016	82 SU / EOG	0.034455	110 HES / SU	0.049279
28 CVE / DVN	0.005135	56 COP / SLB	0.020057	83 OII / NOV	0.035070	111 HAL / SM	0.049931



BACKTESTER

Pair	Trd	PnL	Ann%	Shrp
RIG_NOV	17	17,661	17.81	3.92
OII_RIG	15	15,383	15.52	3.31
CVE_EOG	13	11,615	11.71	2.62
COP_NOV	16	11,432	11.53	3.12
FANG_RIG	14	10,995	11.09	3.36
APA_CVE	16	10,622	10.71	2.82
HAL_RIG	17	10,511	10.60	2.63
RIG_FANG	14	10,183	10.27	3.05
HP_COP	15	9,789	9.87	2.60
APA_EOG	16	9,754	9.84	2.18
FANG_NOV	16	9,523	9.60	2.82
HP_EOG	14	9,432	9.51	2.31
APA_RIG	14	9,071	9.15	2.05
APA_OXY	15	9,030	9.11	3.31
HAL_EOG	14	8,931	9.01	2.38
CVE_APA	13	8,806	8.88	2.33
COP_HP	15	8,595	8.67	2.77
RIG_APA	13	8,574	8.65	2.17
OXY_APA	15	8,495	8.57	3.21
OXY_EOG	14	8,487	8.56	2.39
NOV_COP	15	8,201	8.27	2.25
FANG_EOG	13	8,198	8.27	2.33
SM_COP	15	7,891	7.96	2.33
SM_EOG	12	7,764	7.83	2.22
OII_SM	12	7,514	7.58	2.80
COP_HAL	16	7,386	7.45	2.39
HAL_COP	15	7,372	7.43	2.83
COP_FANG	16	6,987	7.05	3.03
SU_EOG	16	6,871	6.93	3.13
SM_OII	13	6,681	6.74	2.36
APA_SU	13	6,521	6.57	1.95
FANG_APA	14	6,166	6.22	2.25
DVN_CVE	16	6,012	6.06	1.70
MUR_XOM	17	5,968	6.02	1.63
DVN_XOM	17	5,818	5.87	1.97
OXY_HES	12	5,760	5.81	1.42
OII_SU	14	5,739	5.79	1.85
SLB_COP	11	5,717	5.76	2.67
COP_SLB	12	5,622	5.67	2.13
DINO_VLO	14	5,492	5.54	1.47
CNQ_SU	13	5,432	5.48	1.86
FANG_COP	14	5,382	5.43	2.40
OXY_FANG	16	5,359	5.40	2.13
APA_DVN	14	5,173	5.22	1.27

Pair	Trd	PnL	Ann%	Shrp
RIG_DVN	13	5,159	5.20	1.27
FANG_MUR	15	5,141	5.18	1.93
OII_MUR	15	5,140	5.18	1.81
SLB_SU	11	5,132	5.17	1.64
CVE_XOM	16	5,080	5.12	1.41
NOV_FANG	17	5,069	5.11	1.66
HAL_HES	13	4,968	5.01	1.51
OII_HAL	13	4,945	4.99	1.49
MUR_COP	13	4,921	4.96	1.43
MUR_OXY	15	4,723	4.76	1.96
OXY_MUR	14	4,497	4.53	2.15
CNQ_XOM	16	4,448	4.48	1.87
HAL_FANG	12	4,416	4.45	1.40
HAL_XOM	16	4,274	4.31	1.12
MUR_FANG	13	4,250	4.28	1.59
CVE_OXY	14	4,233	4.27	1.52
AR_CTRA	13	4,114	4.15	1.48
DVN_NOV	13	3,978	4.01	2.04
MUR_NOV	11	3,804	3.84	1.37
FANG_OXY	15	3,754	3.78	1.77
OXY_NOV	14	3,737	3.77	1.58
CVX_SU	13	3,729	3.76	1.13
HAL_OII	11	3,689	3.72	1.04
RRC_AR	14	3,651	3.68	1.94
CVE_DVN	14	3,606	3.64	1.01
HAL_SU	13	3,559	3.59	1.28
APA_FANG	11	3,518	3.55	1.33
HAL_MUR	11	3,417	3.45	1.30
COP_XOM	15	3,152	3.18	1.11
AR_RRC	15	3,149	3.17	1.65
APA_NOV	11	2,998	3.02	0.79
HAL_CVX	11	2,845	2.87	0.85
MUR_EOG	11	2,824	2.85	0.89
COP_MUR	12	2,783	2.81	1.25
FANG_SU	12	2,657	2.68	0.77
PAGP_PAA	20	2,648	2.67	3.45
PAA_PAGP	20	2,587	2.61	3.44
SM_SU	13	2,581	2.60	1.30
DVN_SU	13	2,566	2.59	1.17
OII_NOV	12	2,434	2.45	0.52
NOV_OXY	12	2,189	2.21	1.01
SM_HES	14	2,063	2.08	1.02
NOV_APA	12	2,032	2.05	0.56
APA_HES	14	1,987	2.00	0.82

Pair	Trd	PnL	Ann%	Shrp
HAL_OII	11	3,689	3.72	1.04
RRC_AR	14	3,651	3.68	1.94
CVE_DVN	14	3,606	3.64	1.01
HAL_SU	13	3,559	3.59	1.28
APA_FANG	11	3,518	3.55	1.33
HAL_MUR	11	3,417	3.45	1.30
COP_XOM	15	3,152	3.18	1.11
AR_RRC	15	3,149	3.17	1.65
APA_NOV	11	2,998	3.02	0.79
HAL_CVX	11	2,845	2.87	0.85
MUR_EOG	11	2,824	2.85	0.89
COP_MUR	12	2,783	2.81	1.25
FANG_SU	12	2,657	2.68	0.77
PAGP_PAA	20	2,648	2.67	3.45
PAA_PAGP	20	2,587	2.61	3.44
SM_SU	13	2,581	2.60	1.30
DVN_SU	13	2,566	2.59	1.17
OII_NOV	12	2,434	2.45	0.52
NOV_OXY	12	2,189	2.21	1.01
SM_HES	14	2,063	2.08	1.02
NOV_APA	12	2,032	2.05	0.56
APA_HES	14	1,987	2.00	0.82
NOV_HES	14	1,968	1.98	1.15
SM_HAL	13	1,967	1.98	0.59
OXY_SU	13	1,789	1.80	0.93
CVE_CVX	13	1,522	1.53	0.48
CVE_NOV	11	1,396	1.41	0.53
CVE_SU	13	1,374	1.38	0.88
CNQ_HES	13	1,332	1.34	0.76
OII_HES	13	1,311	1.32	0.75
CVE_HES	13	1,296	1.31	0.78
MUR_HES	13	1,104	1.11	0.77
SM_CVX	13	1,056	1.06	0.38
HES_SU	13	869	0.88	0.46
MUR_SU	12	831	0.84	0.62
NOV_MUR	9	747	0.75	0.35
COP_SU	13	398	0.40	1.10
HAL_SM	11	−274	−0.28	−0.06
APA_CVX	11	−442	−0.45	−0.09
OII_OXY	10	−552	−0.56	−0.13
XOM_HES	12	−669	−0.67	0.06
BP_HES	14	−839	−0.85	−0.24
HP_SLB	11	−1,313	−1.32	−0.40
BP_SU	11	−2,115	−2.13	−0.15
MUR_OII	10	−2,159	−2.18	−0.39

Figure 1: Pipeline log. 1,980 tests shrunk to 111 cointegrated pairs

Figure 2: Back-test log ranked by P&L

Figure 1 is an example of our program running. On the left-hand side, 1,980 pairs from one sector are cut down to 111 pairs (NRG). The green values of the p-values represent p-values less than 1%, while the yellow p-values are those that are between 1-5%. Figure 2 on the right-hand side shows the results of the backtester (only tests pairs that are both correlated and cointegrated). We have the pair, the number of trades, the profit and loss, the annualized returns, and the sharpe ratio as our columns.

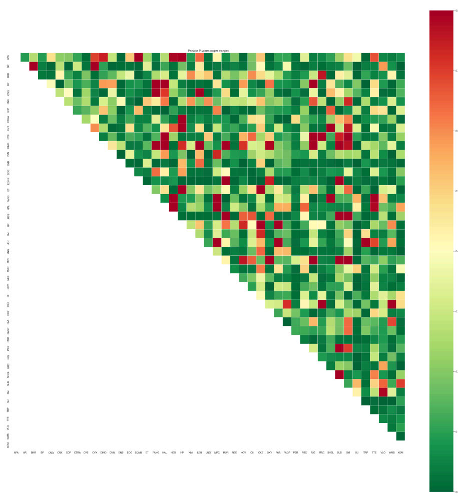


Figure 3: Heat-map of ADF p-values (NRG)

Something to notice from Figure 3 is that from our NRG results is that most of the graph is green. This shows many of the p-values of the possible pairs came close to crossing the threshold of a reasonable p-value.

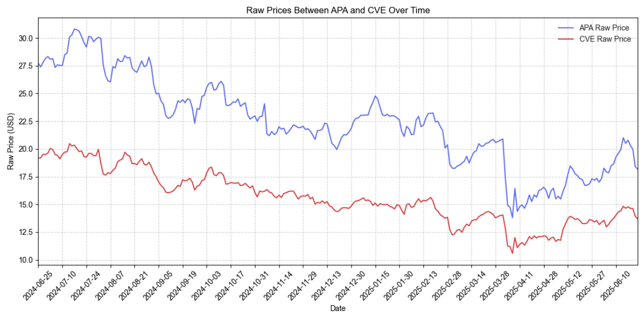


Figure 4: APA vs CVE raw prices through time

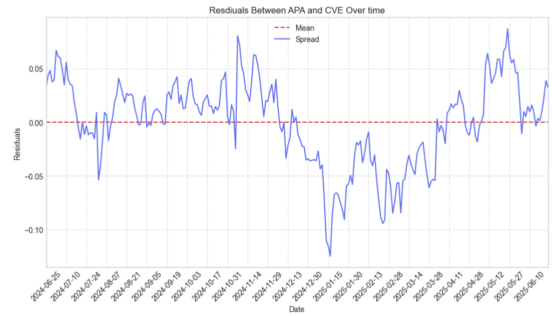


Figure 5: APA-CVE residuals oscillate around zero

In Figure 4, we compare the raw prices of APA and CVE. From this specific pair, correlation is easy to see as these stocks move together. In Figure 5, the residuals are instead plotted. The red line indicates where the mean-line is. This pair of stocks are oscillating around this mean line, which may indicate mean-reversion behavior. Indeed, the high correlation and cointegration from our visualizations are reflected in our results in Figure 2 for profit and loss (PnL).

Discussion

Limitations & Next Steps

Our pairs trading project pushed beyond the requirements, so naturally it carries a few caveats. First, we scraped only one year (≈252 trading days) of daily closes; that limited sample means fewer than 20 trades per pair, inflating metrics like Sharpe and leaving results vulnerable to "lucky" regimes, look-ahead bias, and gaps in the raw data. Second, our screening relied on Pearson correlation and a 5% ADF p-value, simple by design but prone to miss nonlinear links and to admit ~5 % false positives—that's about 100 pairs out of the 1,980 permutations we tested. Third, the back-test assumed frictionless markets, continuous liquidity, and a static hedge ratio β calibrated on the full sample; in reality, bid-ask spreads, borrow fees, slippage, and regime shifts would all clip returns.

We acknowledge these limitations not to take away from the project, but to show we know where our current framework fails and the real-world complexity begins. With more time we would (i) lengthen and clean the price history, (ii) roll the correlation/cointegration screens through expanding windows, (iii) tighten significance to $p < 0.01$ or use Johansen tests, (iv) refresh β with a rolling regression or Kalman filter, and (v) model per-leg transaction costs and capital allocation across concurrent trades. Even so, the project was instrumental in cementing our coding skills and giving us a sandbox to learn the math, statistics, and market mechanics that carry statistical arbitrage. Our future work will simply layer rigor on top of this exploratory foundation.

Conclusion

We built an end-to-end C++ plus Python (*pun-intended)* pipeline that scans thousands of stock pairs, flags those that are cointegrated, and spotlights the most promising pairs-trading setups. For our demo we purposely focused on the Energy sector, and the run already surfaced several encouraging candidates. The results make us cautiously optimistic, and our next step is to tighten the statistical tests and expand coverage.

References & Acknowledgement

Citations

Engle, R. F. & Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. Econometrica https://doi.org/10.2307/1913236
Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs Trading: Performance of a Relative-Value Arbitrage Rule. Review of Financial Studies. https://doi.org/10.1093/rfs/hhj020
Vidyamurthy, G. (2004). Pairs Trading: Quantitative Methods and Analysis. Hoboken, NJ: Wiley Finance.
https://www.wiley.com/en-us/Pairs+Trading%3A+Quantitative+Methods+and+Analysis-p-9780471460671

Acknowledgement

Developed by Francis Padua & Samuel Chun under the mentorship of Dr. Doina Bein.
Funded by CSUF CIC | PCUBED (Center for Inclusive Computing | Pathways, Pipeline, Practice).

Thank you Dr. Bein for this opportunity.