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

Chun Samuel, Padua Francis

Dr. Doina Bein

# Table of Contents

[**Table of Contents**](#_l8tuydaurs16) 1

[**Introduction**](#_7wqw6d4ewxuj) 3

[What is Pairs Trading?](#_9ns7ctnznd) 3

[Origin](#_o834roms7rfc) 3

[How the Strategy Works](#_lz3kh1qhg9l9) 3

[Overview](#_svkzwwgzwfxv) 3

[Objectives](#_qjqsqoe5knst) 4

[Hypothesis](#_8e7jxpa2fcso) 4

[**Methodology**](#_8d96t7sjh6vb) 6

[Collecting Data](#_pa9j86kd6rfy) 6

[Correlation Analysis](#_fe8u6ntv38gp) 6

[Cointegration Testing](#_3vn8wtydw2tq) 7

[**Results**](#_lxswk7vgodo) 8

[**Figures**](#_7q99cwp4v8xi) 8

[**Discussion**](#_u9xlczfrvvli) 10

[Limitations & Next Steps](#_coqmjd5p08wi) 10

[Conclusion](#_f5epn77a07c1) 10

[**References & Acknowledgement**](#_30i2t634209m) 11

[Citations](#_4rbvehatu3q6) 11

[Acknowledgement](#_42za0h1ph68a) 11

# 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.

# 

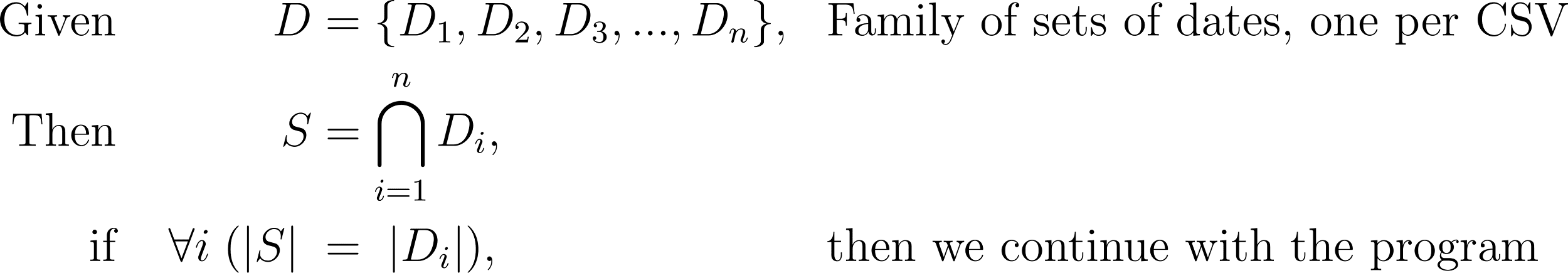
# Methodology

## 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

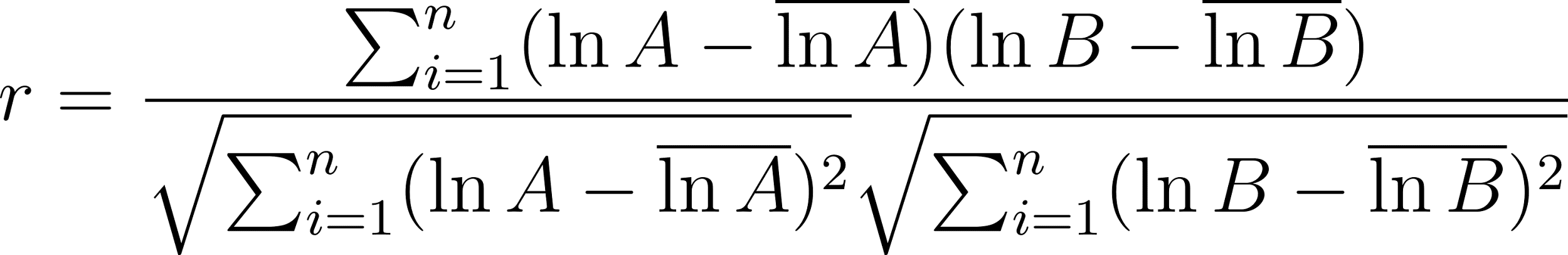
Before we continue with correlation analysis, we use C++ to first guarantee that all date entries are the same in all CSVs. We do so by performing a set intersection on all CSVs for ‘Date’ entries to ensure integrity of data:

[](https://www.codecogs.com/eqnedit.php?latex=%20%5Cbegin%7Balign*%7D%5C%5C%5C%5C%20%5Ctext%7BGiven%7D%5C%3B%26%26%20D%20%26%3D%20%5C%7BD_1%2CD_2%2CD_3%2C...%2CD_n%5C%7D%2C%20%26%26%20%5Ctext%7BFamily%20of%20sets%20of%20dates%2C%20one%20per%20CSV%7D%5C%5C%5C%5C%20%5Ctext%7BThen%7D%5C%3B%26%26%20S%20%26%3D%20%5Cbigcap_%7Bi%3D1%7D%5E%7Bn%7D%20D_i%2C%5C%5C%5C%5C%20%5Ctext%7Bif%7D%20%5C%3B%26%26%20%5Cforall%20i%5C%3B(%5Clvert%20S%20%5Crvert%20%5C%3B%26%3D%5C%3B%20%5Clvert%20D_i%20%5Crvert)%2C%20%26%26%5Ctext%7Bthen%20we%20continue%20with%20the%20program%7D%5C%5C%5C%5C%20%5Cend%7Balign*%7D%20#0)

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:

[](https://www.codecogs.com/eqnedit.php?latex=%20%5Cbegin%7Balign*%7D%5C%5C%5C%5C%20r%20%26%3D%20%5Cfrac%7B%5Csum_%7Bi%3D1%7D%5E%7Bn%7D%20(%5Cln%20A%20-%20%5Coverline%7B%5Cln%20A%7D)(%5Cln%20B%20-%20%5Coverline%7B%5Cln%20B%7D)%7D%7B%5Csqrt%7B%5Csum_%7Bi%3D1%7D%5E%7Bn%7D%20(%5Cln%20A%20-%20%5Coverline%7B%5Cln%20A%7D)%5E2%7D%5Csqrt%7B%5Csum_%7Bi%3D1%7D%5E%7Bn%7D%20(%5Cln%20B%20-%20%5Coverline%7B%5Cln%20B%7D)%5E2%7D%7D%5C%5C%5C%5C%20%5Cend%7Balign*%7D%20#0)

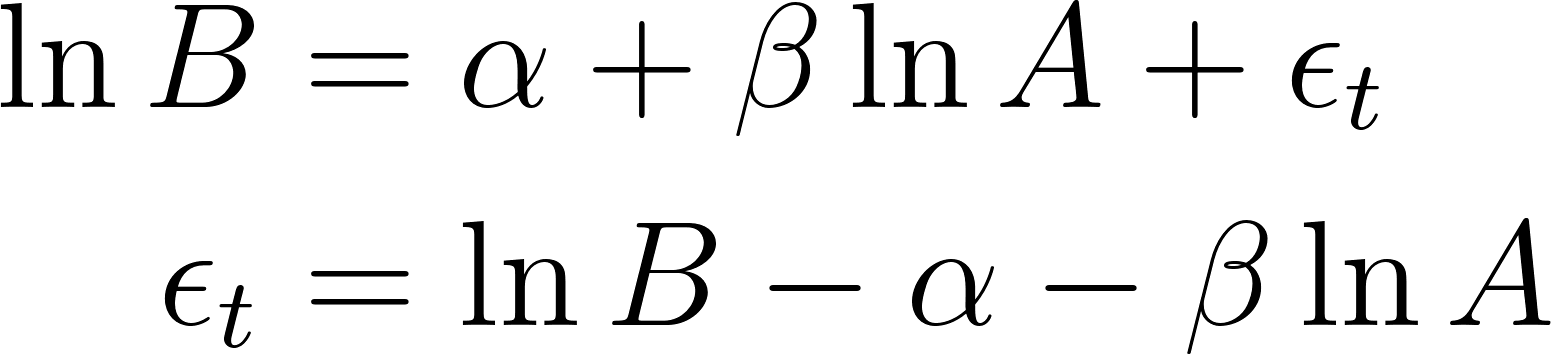
If r > 0.7, then the program concludes that this pair is correlated.

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):

[](https://www.codecogs.com/eqnedit.php?latex=%20%5Cbegin%7Balign*%7D%5C%5C%5C%5C%20%5Cln%20B%20%26%3D%20%5Calpha%20%2B%20%5Cbeta%20%5Cln%20A%20%2B%20%5Cepsilon_t%5C%5C%5C%5C%20%5Cepsilon_t%20%26%3D%20%5Cln%20B%20-%20%5Calpha%20-%20%5Cbeta%20%5Cln%20A%20%5C%5C%5C%5C%20%5Cend%7Balign*%7D%20#0)

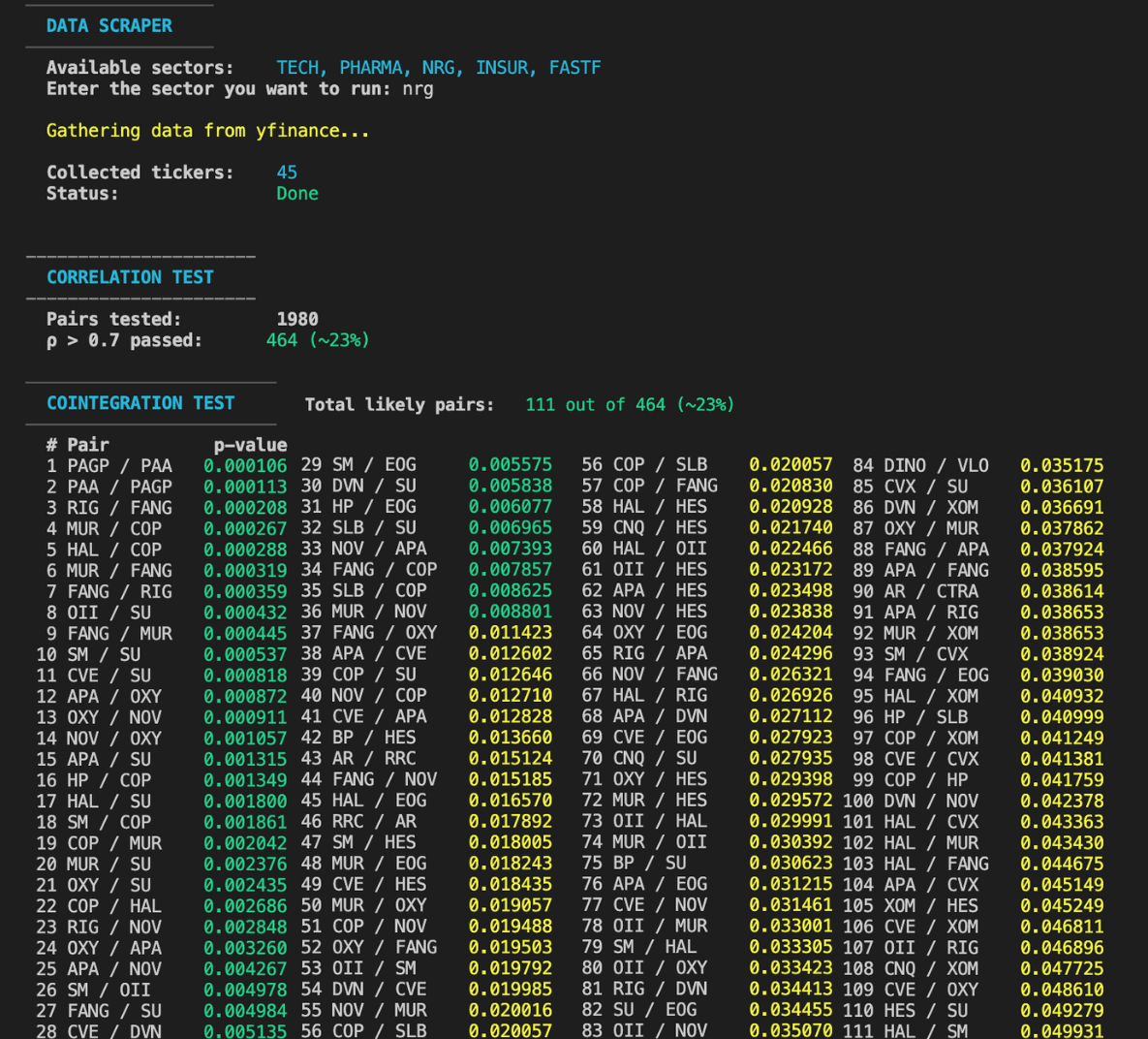
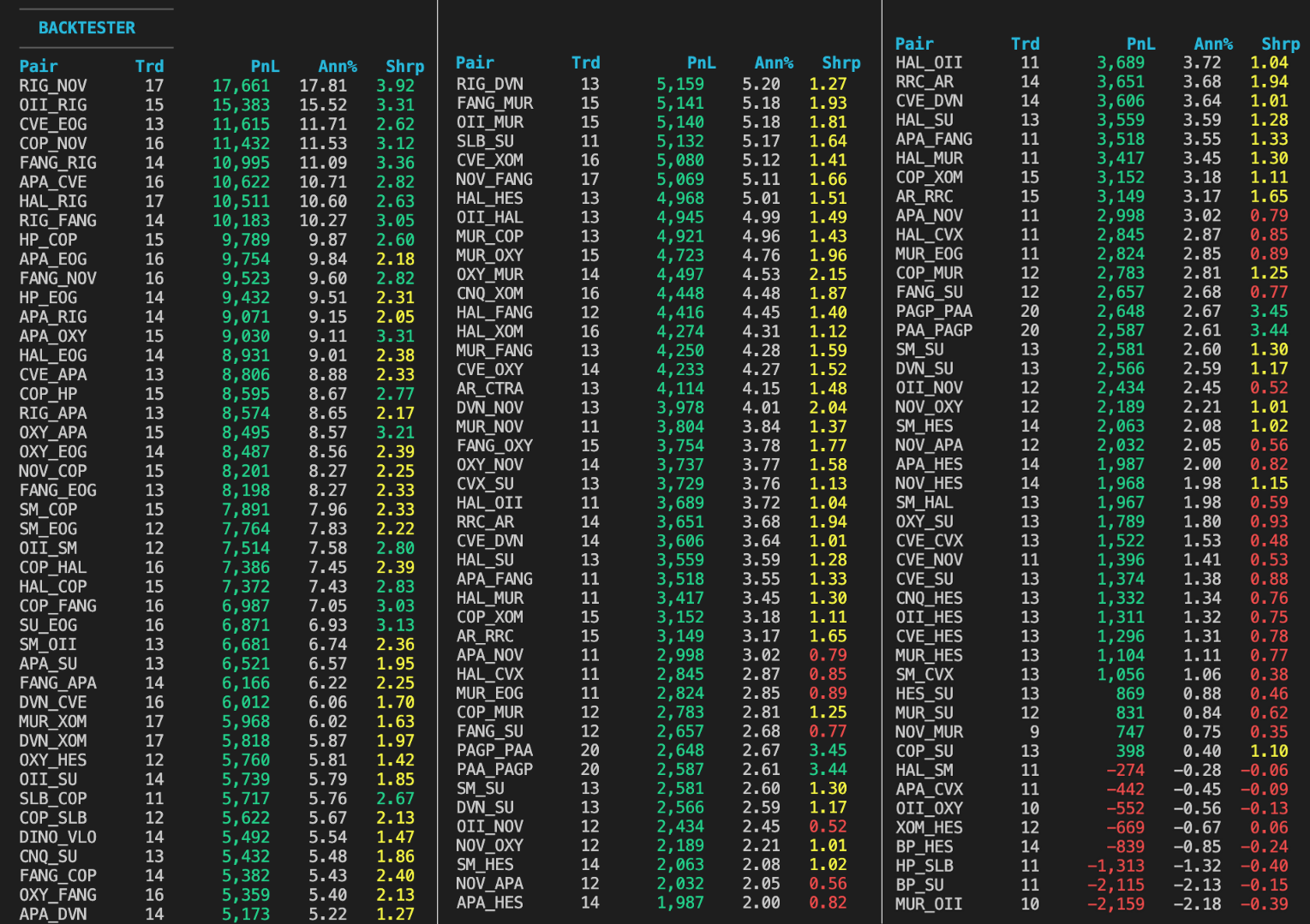
Using these residuals, we may then pass them through the ADF test. ADF confirms stationarity by using statistical hypothesis testing—the null hypothesis of ADF is that the series has a unit root (non-stationary), and the alternative hypothesis stating that our residuals are stationary, returning a p-value. 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.

# 

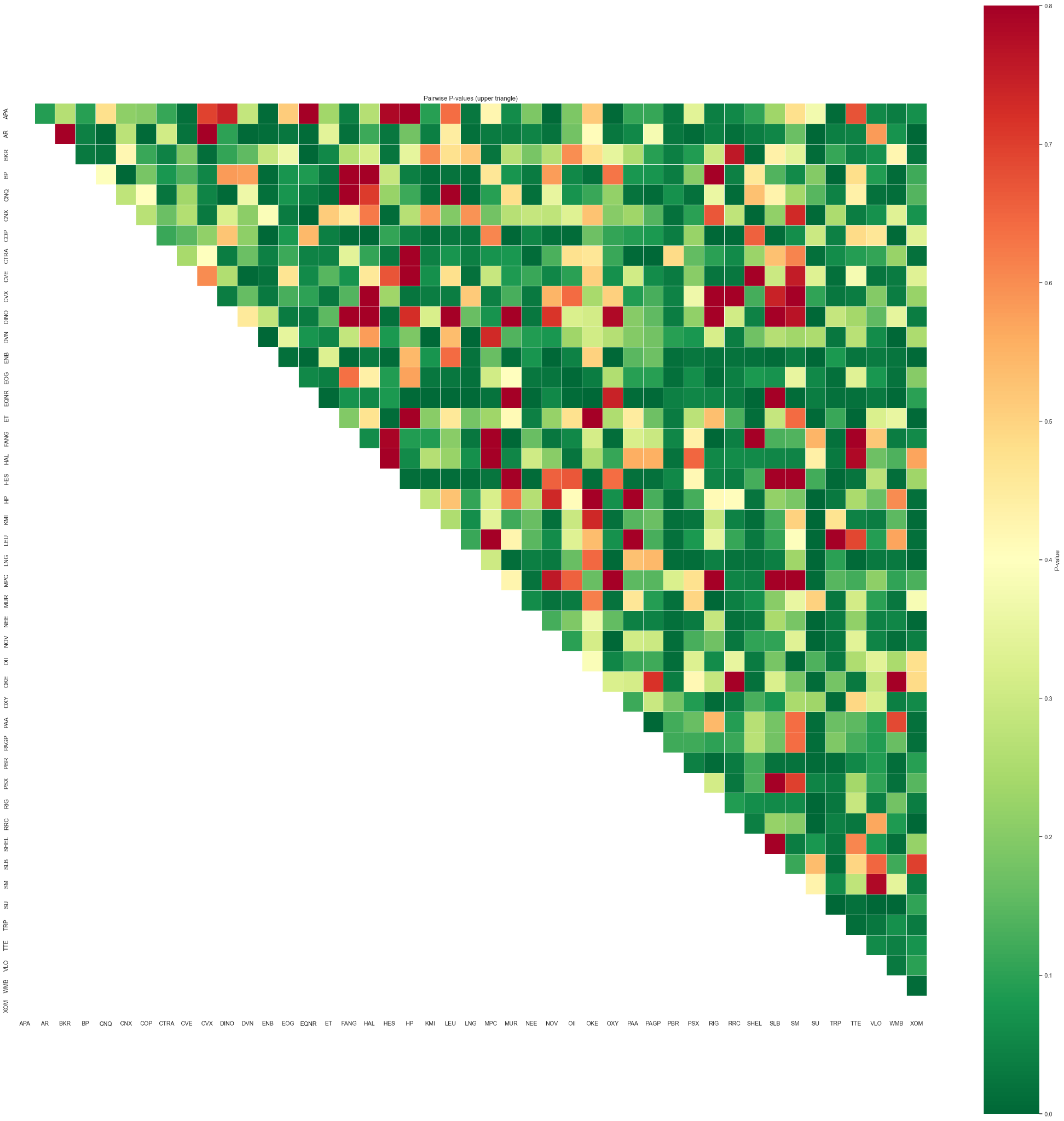
# Results

# Figures

Below are some figures from running our program.

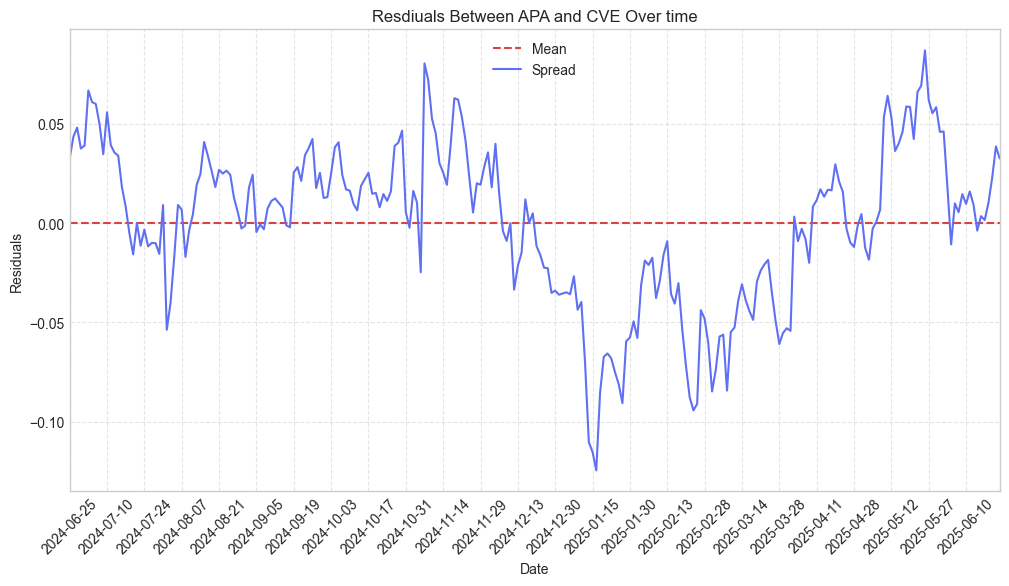
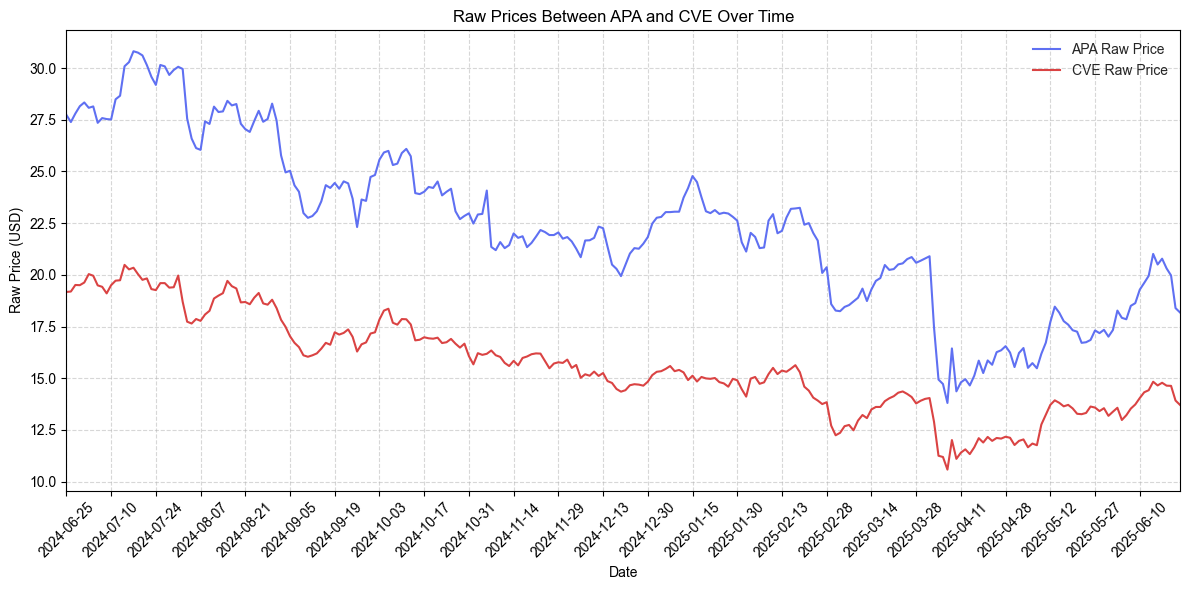
**Figure 1:** Pipeline log. 1,980 tests shrunk to 111 cointegrated pairs **Figure 2:** Back-test log ranked by P&L

*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.