**[Statistical Arbitrage Strategy in R](https://www.quantinsti.com/blog/epat-project-jacques-statistical-arbitrage" \o "Permalink to EPAT Final Project by Jacques Joubert – Statistical Arbitrage Strategy in R)**



Statistical Arbitrage Strategy in R – EPAT Project Work

By [Jacques Joubert](https://za.linkedin.com/in/jacquesjoubert)

On February 29, 2016

From: <https://www.quantinsti.com/blog/epat-project-jacques-statistical-arbitrage>

This article is the final project submitted by the author as a part of his coursework in [Executive Programme in Algorithmic Trading (EPAT)](https://www.quantinsti.com/epat/) at QuantInsti. Do check our Projects [page](https://www.quantinsti.com/category/project-work-epat/) and have a look at what our students are building.

**Background**

For those of you who have been following my blog posts for the last 6 months will know that I have taken part in the [Executive Programme in Algorithmic Trading](https://www.quantinsti.com/courses/epat/) offered by QuantInsti.

It’s been a journey and this article serves as a report on my final project focusing on statistical arbitrage, coded in R. This article is a combination of my class notes and my source code.

I uploaded everything to [GitHub](https://github.com/Jackal08/QuantInsti-Final-Project-Statistical-Arbitrage) in order to welcome readers to contribute, improve, use, or work on this project. It will also form part of my [Open Source Hedge Fund project](http://www.quantsportal.com/open-source-hedge-fund/)on my blog QuantsPortal

I would like to say a special thank you to the team at QuantInsti. Thank you for all the revisions of my final project, for going out of your way to help me learn, and the very high level of client services.

**History of Statistical Arbitrage**

First developed and used in the mid-1980s by Nunzio Tartaglia’s quantitative group at Morgan Stanly.

* Pair Trading is a “contrarian strategy” designed to harness mean-reverting behavior of the pair ratio
* David Shaw, founder of D.E Shaw & Co, left Morgan Stanley and started his own “Quant” trading firm in the late 1980s dealing mainly in pair trading

**What is Pair Trading?**

Statistical arbitrage trading or pairs trading as it is commonly known is defined as trading one financial instrument or a basket of financial instruments – in most cases to create a value neutral basket.

It is the idea that a co-integrated pair is mean reverting in nature. There is a spread between the instruments and the further it deviates from its mean, the greater the probability of a reversal.

Note however that statistical arbitrage is not a risk free strategy. Say for example that you have entered positions for a pair and then the spread picks up a trend rather than mean reverting.

**The Concept**

**Step 1**: Find 2 related securities

Find two securities that are in the same sector / industry, they should have similar market capitalization and average volume traded.

An example of this is Anglo Gold and Harmony Gold.

**Step 2**: Calculate the spread

In the code to follow I used the pair ratio to indicate the spread. It is simply the price of asset A / price asset B.

**Step 3**: Calculate the mean, standard deviation, and z-score of the pair ratio / spread.

**Step 4**: Test for co-integration

In the code to follow I use the Augmented Dicky Fuller Test (ADF Test) to test for co-integration. I set up three tests, each with a different number of observations (120, 90, 60), all three tests have to reject the null hypothesis that the pair is not co-integrated.

**Step 5**: Generate trading signals

Trading signals are based on the z-score, given they pass the test for co-integration. In my project, I used a z-score of 1 as I noticed that other algorithms that I was competing with were using very low parameters. (I would have preferred a z-score of 2, as it better matches the literature, however, it is less profitable)

**Step 6**: Process transactions based on signals

**Step 7**: Reporting

R markdown for my project

**Import packages and set directory**

The first step is always to import the packages needed.

#Imports

require(tseries)

require(urca) #Used for the ADF Test

require(PerformanceAnalytics)

This strategy will be run on shares listed on the Johannesburg Stock Exchange (JSE); because of this I won’t be using the quantmod package to pull data from yahoo finance, instead, I have already gotten and cleaned the data that I stored in a SQL database and moved to CSV files on the Desktop.

I added all the pairs used in the strategy to a folder which I now set to be the working directory.

##Change this to match where you stored the csv files folder name FullList

setwd("~/R/QuantInsti-Final-Project-Statistical-Arbitrage/database/FullList")

**Functions that will be called from within other functions (No user interaction)**

Next: Create all the functions that will be needed. The functions below will be called from within other functions so you don’t need to worry about the arguments.

**AddColumns**

The AddColumns function is used to add columns to the data frame that will be needed to store variables.

#Add Columns to csvDataframe

AddColumns

**PrepareData**

The PrepareData function calculates the pair ratio and the log10 prices of the pair. It also calls the AddColumns function within it.

PrepareData

**GenerateRowValue**

The GenerateRowValue function Calculates the mean, standard deviation and the z-score for a given row in the data frame.

#Calculate mean, stdDev, and z-score for the given Row [end]

GenerateRowValue

**GenerateSignal**

The GenerateSignal function creates a long, short, or close signal based on the z-score. You can manually change the z-score. I have set it to 1 and -1 for entry signals and any z-score between 0.5 and -0.5 will create a close/exit signal.

GenerateSignal trigger)

csvData$signal[counter] -close)

csvData$signal[counter]

**GenerateTransactions**

The GenerateTransactions function is responsible for setting the entry and exit prices for the respective long and short positions needed to create a pair.

Note: QuantInsti taught us a very specific way of backtesting a trading strategy. They used excel to teach strategies and when I coded this strategy I used a large part of the excel methodology.

Going forward, however, I would explore other ways of storing variables. One of the great things about this method is that you can pull the entire data frame and analyse why a trade was made and all the details pertaining to it.

#Transactions based on trade signal

#Following the framework set out initially by QuantInsti (Note: this can be coded better)

GenerateTransactions

**GetReturnsDaily**

GetReturnsDaily calculates the daily returns on each position and then calculates the total returns and adds slippage.

#Calculate daily returns generated

#Add implementation shortfall / slippage at close of trade

GetReturnsDaily 0){csvData$LongReturn[end] 0){csvData$ShortReturn[end]

**GenerateReports**

The next two arguments are used to generate reports. A report includes the following: Charting: 1. An Equity curve 2. Drawdown curve 3. Daily returns bar chart

Statistics: 1. Annual Returns 2. Annualized Sharpe Ratio 3. Maximum Drawdown

Table: 1. Top 5 drawdowns and their duration

Note: If you have some extra time then you can further break this function down into smaller functions inorder to reduce the lines of code and improve usability. Less code = Less Bugs

#Returns an equity curve, annualized return, annualized sharpe ratio, and max drawdown

GenerateReport 0){

positiveTrades

**Functions that the user will pass parameters to**

The next two functions are the only functions that the user should fiddle with.

**BacktestPair**

BacktestPair is used when you want to run a backtest on a trading pair (the pair is passed in via the CSV file)

Functions arguments:

* pairData = the CSV file date
* mean = the number of observations used to calculate the mean of the spread.
* slippage = the amount of basis points that act as brokerage as well as slippage
* adfTest = a boolean value – if the backtest should test for co-integration
* criticalValue = Critical Value used in the ADF Test to test for co-integration
* generateReport = a boolean value – if a report must be generated

#The function that will be called by the user to backtest a pair

BacktestPair 130){

begin = mean){

#Generate Row values

pairData

**BacktestPortfolio**

BacktestPortfolio accepts a vector of CSV files and then generates an equally weighted portfolio.

Functions arguments:

* names = an attomic vector of CSV file names, example: c(‘DsyLib.csv’, ‘OldSanlam.csv’)
* mean = the number of observations used to calculate the mean of the spread.
* leverage = how much leverage you want to apply to the portfolio

#An equally weighted portfolio of shares

BacktestPortfolio

Running Backtests

Now we can start testing strategies using our code.

**Pure arbitrage on the JSE**

When starting this project the main focus was on using statistical arbitrage to find pairs that were co-integrated and then to trade those, however, I very quickly realized that the same code could be used to trade shares that had both its primary listing as well as access to its secondary listing on the same exchange.

If both listings are found on the same exchange, it opens the door for a pure arbitrage strategy due to both listings referring to the same asset. Therefore you don’t need to test for co-integration.

There are two very obvious examples on the JSE.

**First Example Investec:**

Primary = Investec Ltd : Secondary = Investec PLC

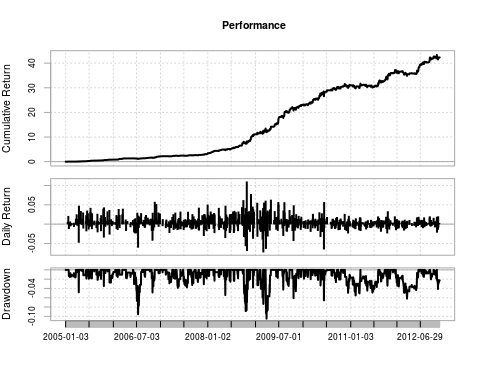
**Investec In-Sample Test (2005-01-01 – 2012-11-23)**

Test the following parameters

* The Investec ltd / plc pair
* mean = 35
* Set adfTest = F (Dont test for co-integration)
* Leverage of x3

#Investec

leverage



## [1] "Annual Returns: 0.619853087807437"

## [1] "Annualized Sharpe: 3.29778431709924"

## [1] "Max Drawdown: 0.105016628973292"

## From Trough To Depth Length To Trough Recovery

## 1 2009-03-19 2009-03-25 2009-05-04 -0.1050 28 5 23

## 2 2006-06-08 2006-07-13 2006-08-14 -0.0955 46 25 21

## 3 2008-10-03 2008-10-17 2008-10-24 -0.0887 16 11 5

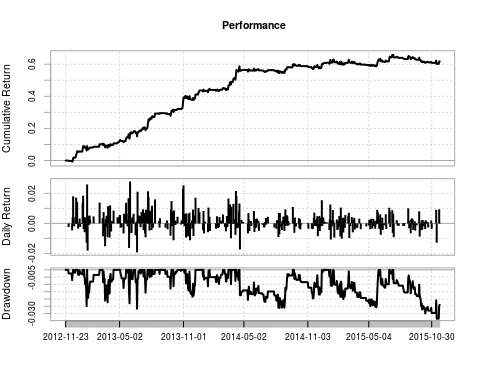
## 4 2009-03-02 2009-03-02 2009-03-06 -0.0733 5 1 4

## 5 2008-10-27 2008-10-27 2008-11-05 -0.0697 8 1 7

**Investec Out-of-Sample Test (2012-11-23 – 2015-11-23)**

Note: if you increase the slippage, you will very quickly kiss profits goodbye.

GenerateReport.xts(investec.returns, startDate = '2012-11-23', endDate = '2015-11-23')



## [1] "Annual Returns: 0.1754103210963"

## [1] "Annualized Sharpe: 2.20385429706265"

## [1] "Max Drawdown: 0.0335642102186873"

## From Trough To Depth Length To Trough Recovery

## 1 2015-07-10 2015-11-13 -0.0336 96 89 NA

## 2 2013-06-18 2013-06-21 2013-07-01 -0.0267 10 4 6

## 3 2014-04-16 2014-08-13 2014-09-19 -0.0262 107 80 27

## 4 2015-01-20 2015-05-25 2015-06-01 -0.0258 91 86 5

## 5 2013-01-18 2013-01-24 2013-01-25 -0.0249 6 5 1

**Second Example Mondi:**

Primary = Mondi Ltd : Secondary = Mondi PLC

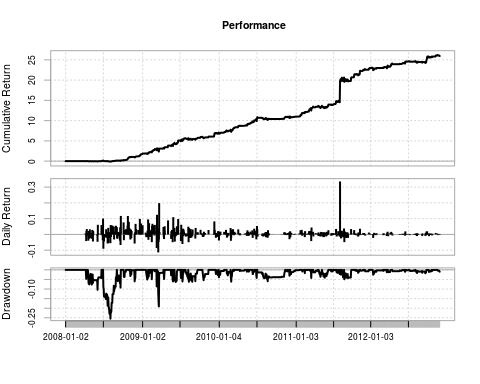
**Mondi In-Sample Test (2008-01-01 – 2012-11-23)**

Test the following parameters

* The Mondi ltd / plc pair
* mean = 35
* Set adfTest = F (Dont test for co-integration)
* Leverage of x3

data mondi

mondi.returns



## [1] "Annual Returns: 0.973552250431717"

## [1] "Annualized Sharpe: 2.88672185296756"

## [1] "Max Drawdown: 0.254688711989788"

## From Trough To Depth Length To Trough Recovery

## 1 2008-07-01 2008-08-01 2008-09-01 -0.2547 45 24 21

## 2 2009-03-11 2009-03-18 2009-04-08 -0.1906 21 6 15

## 3 2008-04-16 2008-06-03 2008-06-23 -0.1040 45 32 13

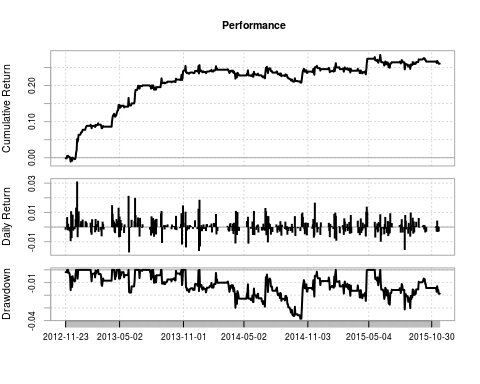
## 4 2008-09-02 2008-09-17 2008-09-18 -0.0926 13 12 1

## 5 2009-03-09 2009-03-09 2009-03-10 -0.0864 2 1 1

**Mondi Out-of-Sample Test (2012-11-23 – 2015-11-23)**

Note: In all of my testing I found that the further down the timeline my data was, the harder it was to make profits on the end of day data. I tested this same strategy on intraday data and it has a higher return profile.

GenerateReport.xts(mondi.returns, startDate = '2012-11-23', endDate = '2015-11-23')



## [1] "Annual Returns: 0.0809094579019469"

## [1] "Annualized Sharpe: 1.25785312960412"

## [1] "Max Drawdown: 0.0385234269750542"

## From Trough To Depth Length To Trough Recovery

## 1 2013-12-19 2014-10-13 2015-01-26 -0.0385 273 202 71

## 2 2015-06-05 2015-08-14 -0.0313 120 49 NA

## 3 2015-01-27 2015-04-22 2015-04-28 -0.0245 63 60 3

## 4 2013-05-29 2013-05-30 2013-06-14 -0.0179 13 2 11

## 5 2013-11-08 2013-11-18 2013-12-18 -0.0175 28 7 21

Next, we will look at a pair trading strategy.

Typically a pair consists of 2 shares that:

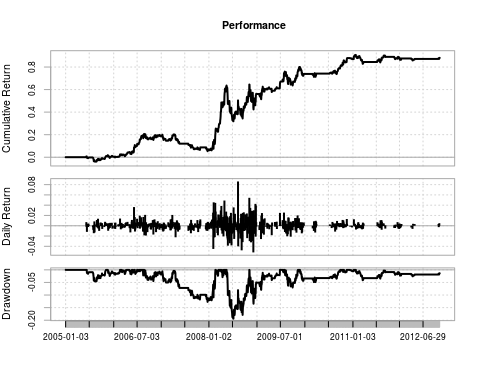
* Share a market sector
* Have a similar market cap
* Similar business model and clients
* Are co-integrated

In all of the portfolios below I use 3x leverage

**Construction Portfolio**

**In-sample test (2005-01-01 – 2012-11-01)**

names



[1] "Annual Returns: 0.0848959306632411"

## [1] "Annualized Sharpe: 0.733688101181479"

## [1] "Max Drawdown: 0.193914686702112"

## From Trough To Depth Length To Trough Recovery

## 1 2008-05-19 2008-07-08 2008-11-03 -0.1939 119 36 83

## 2 2008-11-04 2008-12-03 2009-06-29 -0.1345 160 22 138

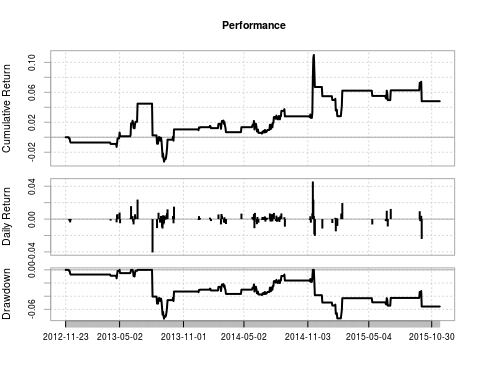
## 3 2006-08-25 2007-12-19 2008-02-19 -0.1272 372 331 41

## 4 2009-08-04 2009-10-01 2009-11-10 -0.0701 69 41 28

## 5 2009-11-25 2010-03-10 2010-09-29 -0.0486 211 73 138

**Out-of-sample test (2012-11-23 – 2015-11-23)**

GenerateReport.xts(ReturnSeries, startDate = '2012-11-23', endDate = '2015-11-23')



## [1] "Annual Returns: 0.0159094762396512"

## [1] "Annualized Sharpe: 0.268766025866724"

## [1] "Max Drawdown: 0.0741426720423424"

## From Trough To Depth Length To Trough Recovery

## 1 2013-08-05 2013-09-06 2014-11-17 -0.0741 322 24 298

## 2 2014-11-20 2015-01-29 -0.0737 253 47 NA

## 3 2012-11-30 2013-04-23 2013-05-02 -0.0129 102 96 6

## 4 2013-06-10 2013-06-13 2013-06-24 -0.0100 10 4 6

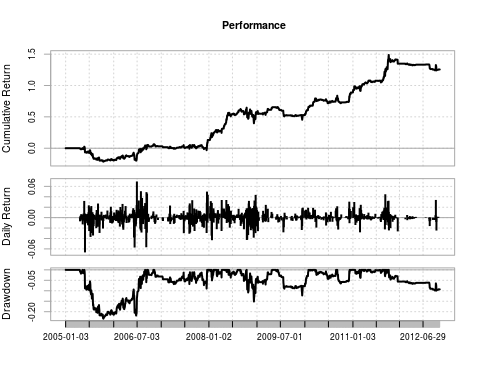
## 5 2013-05-03 2013-05-03 2013-06-04 -0.0050 23 1 22

**[](https://www.quantinsti.com/epat/)**

**Insurance Portfolio**

**In-sample test (2005-01-01 – 2012-11-01)**

names



## [1] "Annual Returns: 0.110600985165525"

## [1] "Annualized Sharpe: 0.791920916349154"

## [1] "Max Drawdown: 0.233251846760865"

## From Trough To Depth Length To Trough Recovery

## 1 2005-05-26 2005-10-14 2006-08-31 -0.2333 318 100 218

## 2 2008-10-15 2008-12-05 2009-04-30 -0.1513 134 38 96

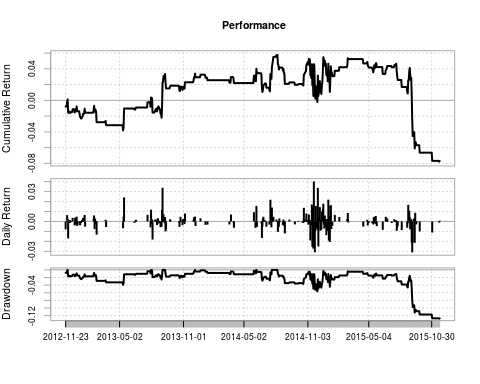
## 3 2009-06-10 2009-12-10 2010-01-29 -0.1223 162 129 33

## 4 2011-10-04 2012-10-09 -0.0991 267 249 NA

## 5 2006-11-08 2007-12-11 2007-12-14 -0.0894 277 274 3

**Out-of-sample test (2012-11-23 – 2015-11-23)**

GenerateReport.xts(ReturnSeries, startDate = '2012-11-23', endDate = '2015-11-23')



## [1] "Annual Returns: -0.0265926093350092"

## [1] "Annualized Sharpe: -0.319582293135835"

## [1] "Max Drawdown: 0.128061204573991"

## From Trough To Depth Length To Trough Recovery

## 1 2014-08-08 2015-11-20 -0.1281 326 324 NA

## 2 2012-11-28 2013-05-13 2013-07-31 -0.0393 167 111 56

## 3 2014-06-10 2014-06-26 2014-07-23 -0.0284 31 12 19

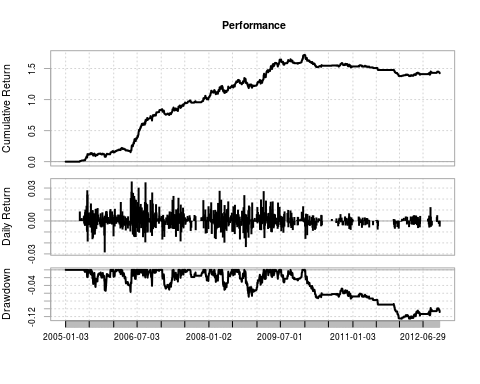
## 4 2013-08-01 2013-08-30 2013-09-03 -0.0255 23 21 2

## 5 2013-09-11 2013-10-22 2013-12-04 -0.0209 60 29 31

**General Retail Portfolio**

**In-sample test (2005-01-01 – 2012-11-01)**

names



## [1] "Annual Returns: 0.120956981644048"

## [1] "Annualized Sharpe: 1.4694780839876"

## [1] "Max Drawdown: 0.125406256082082"

## From Trough To Depth Length To Trough Recovery

## 1 2010-01-05 2012-01-17 -0.1254 705 504 NA

## 2 2008-09-29 2008-10-29 2009-02-20 -0.0690 101 23 78

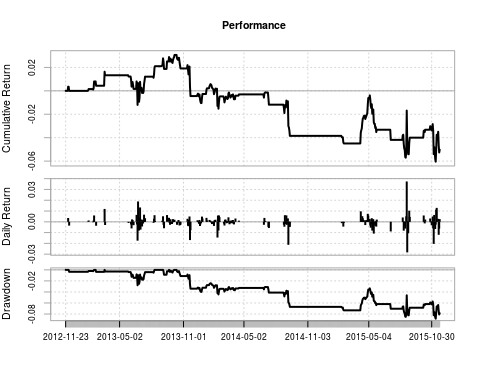
## 3 2006-03-06 2006-05-15 2006-05-23 -0.0568 52 46 6

## 4 2005-07-18 2005-11-01 2005-12-06 -0.0538 101 76 25

## 5 2008-04-11 2008-04-29 2008-06-26 -0.0512 51 12 39

**Out-of-sample test (2012-11-23 – 2015-11-23)**

GenerateReport.xts(ReturnSeries, startDate = '2012-11-23', endDate = '2015-11-23')



[1] "Annual Returns: -0.0171898953593881"

## [1] "Annualized Sharpe: -0.336265418351652"

## [1] "Max Drawdown: 0.0884145115767888"

## From Trough To Depth Length To Trough Recovery

## 1 2013-10-15 2015-11-11 -0.0884 528 519 NA

## 2 2013-03-18 2013-06-24 2013-08-12 -0.0279 100 66 34

## 3 2013-09-05 2013-09-06 2013-09-20 -0.0088 12 2 10

## 4 2013-09-23 2013-10-02 2013-10-08 -0.0049 11 7 4

## 5 2013-02-20 2013-02-20 2013-03-15 -0.0037 18 1 17

**Conclusion:**

At the end of all my testing, and trust me – there is a lot more testing I did than what is in this report, I came to the conclusion that the Pure Arbitrage Strategy has great hope in being used as a strategy using real money, but the Pair Trading Strategy on portfolios of stocks in a given sector is strained and not likely to be used in production in its current form.

There are many things that I think could be added to improve the performance. Going forward I will investigate using Kalman filters.

**More on the Pure Arbitrage Trading Strategy:**

I have only found two shares that have duel listings on the same exchange; this means that we can’t allocate large sums of money to the strategy as it will have a high market impact, however, we could use multiple exchanges and increase the number of shares used.

**More on the Pair Trading Strategy:**

1. The number of observations used in the ADF Tests is large to blame. The problem is that a test for co-integration has to be done in order to make a claim for statistical arbitrage, however by using 120, 90, and 60 as parameters to the three tests, it is very difficult to find pairs that match the criteria and that will continue in this form for the near future. (Kalman filtering may be useful here)
2. I haven’t spent a lot of time changing the different parameters like the number of observations in the mean calculation. (This requires further exploration)
3. From the above sector portfolios, we can see that the early years are very profitable but the further down the timeline we go, the lower returns get. I have spoken to a few people in the industry as well as my friends doing stat arb projects at the University of Cape Town, the local lore has it that in 2009 Goldman switched on their stat arb package, in regards to the JSE listed securities.
4. The same is noticed with other portfolios that I didn’t include in this report but is in the R Code file.
5. I believe that this is due to large institutions using the same bread and butter strategy. You will note (if you spend enough time testing all the strategies) that in 2009 there seems to be a sudden shift in the data to lower returns.
6. I feel that the end of day data I am using is limiting me and if I were to test the strategy on intraday data then profits would be higher. (I ran one test on intraday data on Mondi and the results were much higher, but I am still to test it on sector portfolios)
7. This is one of the simpler statistical arbitrage strategies and I believe that if we were to improve the way we calculate the spread and change some of the entry and exit rules, the strategy would become more profitable.

If you made it to the end of this article, I thank you and hope that it added some value. This is the first time that I am using Github, so I am looking forward to seeing if there are any new contributors to the project.

Github repository: <https://github.com/Jackal08/QuantInsti-Final-Project-Statistical-Arbitrage>

Next Step

Read about other strategies in this article on [Algorithmic Trading Strategy Paradigms](https://www.quantinsti.com/blog/algorithmic-trading-strategies/). If you want to learn Algorithmic Trading, then [click here](https://www.quantinsti.com/epat).

**Update**

*We have noticed that some users are facing challenges while downloading the market data from Yahoo and Google Finance platforms. In case you are looking for an alternative source for market data, you can use [Quandl](https://www.quandl.com/collections/markets" \t "_blank) for the same.*