Pairs Trading

Big Data Project

Varun Seshu

PES University
Bangalore, India
PES2201800074
varunseshu@gmail.com

Hritik shanbhag

PES University
Bangalore,India
PES2201800082
hritik.24shanbhag@gmail.com

Shashwath Kumar S

PES University
Bangalore, India
PES2201800623
shashwath457@gmail.com

Abstract

This project endeavours to simulate pair trading and find correlated and co-integrated pairs in the indian stock market, generate orders based on the pairs identified and evaluate the trades made on the pairs. Since the number of pairs generated are large (~25 million) we need to use BIg Data technologies to achieve this objective.

You can access the repository of the Code for the project at this link:
https://github.com/Varun487/BigData_Pair_Trading

Introduction

Pairs trading is simple: find a pair of stocks which track each other, and when they diverge, buy the lower one and short the high. If they converge again, you pocket the difference.

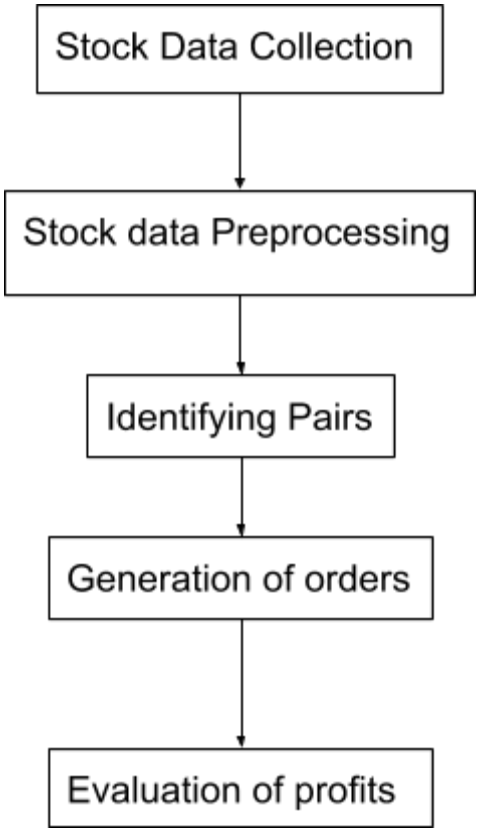
Pairs Trading can be called a Mean Reversion Strategy - A strategy where we one bets that the prices will eventually revert to their historical trends if there is a significant deviation. The most basic or the first step is to find pairs that are highly cointegrated or correlation.The best way to do this is to start with stocks that might be cointegrated (based on some common properties maybe?) and perform a statistical test. Usually the businesses that are in the same industry or sub-sector show good correlation. We then use **PYSPARK** to evaluate all possible pairs in each sector of the stock market (~25 million pairs) to find good pairs and explot this strategy.

After finding a pair, 'we can expect the ratio or difference in prices (also called the spread) of these two pairs to remain constant with time. However, sometimes there might be a divergence in the spread between these two pairs caused by temporary supply or demand changes, large buy or sell orders for one stock, reaction for important news about one of the companies etc. In this scenario,if one stock moves up while the other moves down relative to each other. If you expect this divergence to revert back to normal with time, you can make a pairs trade.

When we observe a temporary divergence,the basic idea is to sell the outperforming stock i.e the stock that went high and to buy the underperforming stock-the stock that went down and assume that the spread between the two stocks would eventually converge by either the outperforming stock moving back down or underperforming stock moving back up or both — whatever the scenario we will make the money. However if both the stocks move up or move down together without changing the spread between them, no money is lost or gained.Therefore pairs trading is a market neutral trading strategy enabling traders to profit from any of the market conditions: uptrend, downtrend, or sideways movement.

Project Workflow

This is a broad outline of the steps taken in this project.



In **Stock data collection**, we collect historical candlestick data up to 20 years for all stocks actively trading in the stock market.

In **Stock data preprocessing**, we clean the data - handle missing data, crop the datasets, drop columns, etc. - and prepare it for the next step.

In **Identifying Pairs**, we use the cleaned data and generate and identify correlated and cointegrated pairs for each sector.We filter to only include pairs that fit our criteria of correlation and cointegration.

In **Generation of Orders**, After we have correlated and cointegrated pairs, we generate orders using our pair trading algorithm (mentioned in the ‘Trading Methodology’ section) over the period where they are correlated and cointegrated.

In **Evaluation of profits**, we then evaluate the profits/losses made by each order made on the pair.

Trading Methodology

In trading methodology, we outline the parameters we use to generate orders and perform the experiment.

Assumptions based on pair trading:

1) Pair trading is a mean reverting strategy, where we make a bet expecting the spread between the pair of stocks to revert back to the mean spread from its current position. Thus, when the spread increases, we bet that it decreases back and reverts to the mean. Vice Versa when the Spread decreases.

2) Thus we take a Short position when the Spread increases beyond a threshold and a Long position when the Spread decreases beyond a threshold.

Criteria to place orders and trade:

1) **Capital**: The Capital we are using to perform the trades per stock is Rs. 10,00,000.

2) **Risk**: We are allowing a maximum risk of 2% of the total capital per trade, i.e, Rs. 20000.

3) **Finding pairs**: For each sector in the market, we create pairs of stocks and find 90 day periods where they are correlated and cointegrated. We choose stock pairs with correlation coefficient $r > 0.85$ and with p-value < 0.05 in the cointegration test to see if the correlation calculated is statistically significant.We then choose the 4 best stock pairs according to correlation and cointegration value for the rest of the experiment.

4) **Spread**: We then calculate the spread (Refer to Appendix) for the chosen stock pairs.

5) **z-score**: We calculate the z-score of the spreads, using which we generate orders to place based on thresholds below.

6) **Open position**: We open a Short position on the pair if the z-score of the spread > 1.5, i.e, Short position on stock 1 and Long position on stock 2.We open a Long position on the pair if the z-score of the spread < -1.5, i.e, Long position on stock 1 and Short position on stock 2.

7) **Close Position**: We close a position on the basis of 2 criteria.

8) **Take Profit**: We consolidate our position in order to take the profits.

9) **Stop Loss**: We consolidate our position in order to cut losses in our trade.

10) **Long position**: Take profit criteria - z-score ≥ 0. Stop Loss criteria - z-score ≥ 3.

11) **Short position**: Take profit criteria - z-score ≤ 0. Stop Loss criteria - z-score≤ -3.

PROJECT CODE

The full code and visualizations are present in the Git Repo at:
https://github.com/Varun487/BigData_Pair_Trading

Data Collection

We have scraped Indian stock market daily candlestick data for a period of 20 years on all actively trading stocks from the Yahoo Finance API.

The code for scraping ticker of all stocks in NSE and BSE:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/Collection/list_of_nse_companies.py

We have collected data for approximately **5000 stocks**, each with Columns -

Sample data of 1 stock:

Date	High	Low	Open	Close	Volume	Adj Close
18-07-2000	125	125	125	125	100	84.24127
19-07-2000	123.75	117.5	117.5	120.625	200	81.29282
20-07-2000	120.625	120.625	120.625	120.625	0	81.29282
21-07-2000	120.625	120.625	120.625	120.625	0	81.29282
24-07-2000	120.625	120.625	120.625	120.625	0	81.29282
25-07-2000	120.625	120.625	120.625	120.625	0	81.29282
26-07-2000	120.625	120.625	120.625	120.625	0	81.29282
27-07-2000	120.625	120.625	120.625	120.625	0	81.29282
28-07-2000	120.625	120.625	120.625	120.625	0	81.29282
31-07-2000	120.625	120.625	120.625	120.625	0	81.29282
01-08-2000	120.625	120.625	120.625	120.625	0	81.29282
02-08-2000	120.625	120.625	120.625	120.625	0	81.29282
03-08-2000	120.625	120.625	120.625	120.625	0	81.29282

Date - Date in format yyyy-mm-dd
High - High price of the day
Low - Low price of the day
Open - Open price of the day
Close - Close price of the day
Adjusted Close - Close price adjusted for stock splits
Volume - Number of trades of the stock in a day

We also have a Sector-wise stock ticker csv file which helps us to find pairs of companies within the same sector.

	A	B	C	D	E	F	G	H	I	J
1	Security Code	Issuer Name	Security Id	Security Name	Status	Group	Face	ISIN No	Industry	Instrument
2	500002		ABB	ABB India Limited	Active	B	2	INE117A01022	Heavy Electrical Equipment	Equity
3	500003		AEGISLOG	AEGIS LOGISTICS LTD.	Active	A	1	INE208C01025	Oil Marketing & Distribution	Equity
4	500004		TPAEC	TORRENT POWER AEC LTD.	Delisted	B	10	INE424A01014		Equity
5	500005		AKARLAMI	AKAR LAMINATORS LTD.	Delisted	XD	10	INE984C01013	Iron & Steel Products	Equity
6	500006		ALPHADR	ALPHA DRUG INDIA LTD.	Delisted	B	10	INE256B01026		Equity
7	500008		AMARAJAB	AMARA RAJA BATTERIES LTD.	Active	A	1	INE885A01032	Auto Parts & Equipment	Equity

Lastly, we have 2 csvs, NSETICKERS.csv and BSETICKERS.csv, which consists of all actively traded stocks in the Stock exchanges. We use this data to automate collection of stocks historical data using the Yahoo Finance API.

The code for creating the companies data is present here:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/Collection/stock_candle_data_and_volume.py

Data Preprocessing:

1. **Cleaning Data**

We are first extracting all the csvs from its company folders.
The code for extracting the csvs:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/Preprocessing_dataset/extract_csvs.py

We are dropping rows of the datasets which have missing data. We are not Interpolating the missing data as filling in the data for missing days may lead to discrepancies and might affect the mean reverting nature of larger stocks and increase volatility.

The code for handling missing data:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/Preprocessing_dataset/handle_missing_data.py

We then delete datasets which have data for less than 3 years as we would not have significant correlations and very less data would be available for training and testing models. We then crop the remaining datasets to be between the years 2017-2020 as to reduce compute time for the models and have enough data to find significant correlations. We have chosen not to take 2020

stock data as due to the pandemic, the volatility of the stocks have increased and this may lead to skewed overall results.

The code for cropping the dataset:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/Preprocessing_dataset/crop_data_in_range.py

2. **Generate pairs**

To Identify correlated and cointegrated pairs, we follow the algorithm:

```
For each sector:
  For each stock in sector:
    For each 90 day window in stock data:
      Calculate correlation on close price
      Calculate p-value of cointegration test on close price
      If correlation > 0.85 and p-value < 0.05
         Record pair data and time period
      Else
           move window by 15 days
Store all recorded pair data in csvs.
```

In the above algorithm, all stock data is present in **PYSPARK** dataframes.

3. **Calculate spread and z-score**

For the pairs that we have stored, calculate the Spread of the pair by simply subtracting the prices of the stocks. We then calculate the z-score of the spread of the pair.

The code for identifying and generating pairs:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/generate_pairs_orders_profit/identify_pairs.py

4. **Generate trade orders**

For each pair, we go through the z-score of the spread calculated and open or close positions depending on the criteria mentioned in the Trading Methodology.

We then label each day as:

LONG - Denotes buy the first stock and sell the second.

SHORT - Denotes sell the first stock and buy the second.

FLAT - Denotes no order to be placed on that day.

GET_OUT_OF_POSITION - denotes to cash in on all previous orders on that date and have no outstanding LONG or SHORT positions as of that date.

The code to generate trade orders:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/generate_pairs_orders_profit/generate_orders.py

5. **Evaluating orders and Calculating Profits:**

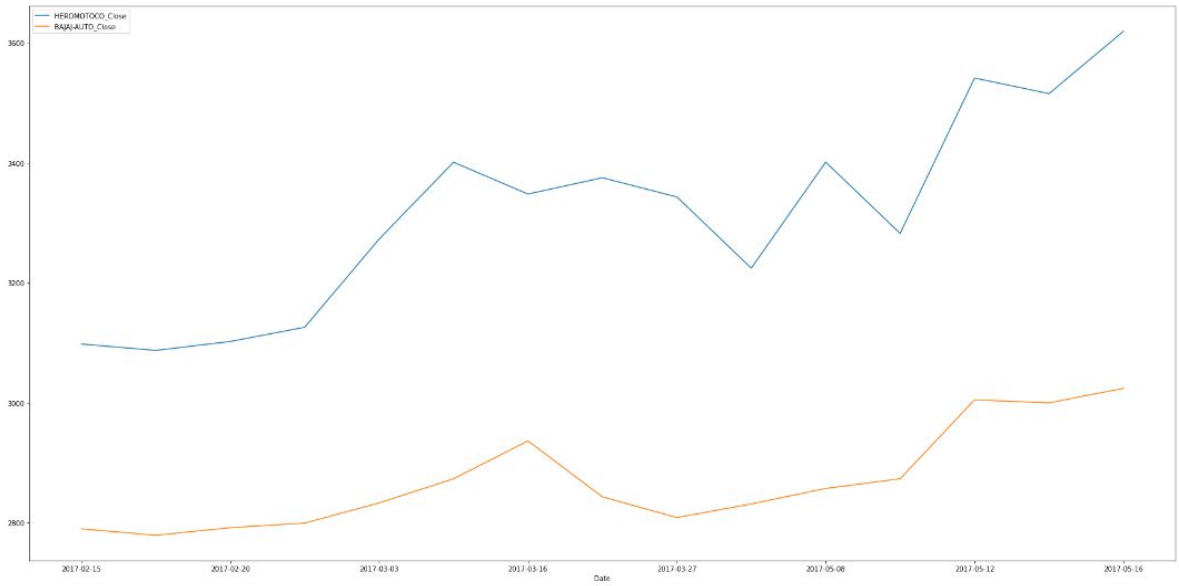
Using capital and risk as mentioned in Trading Methodology above, we evaluate the orders for each pair and get the profit earned per trade.

The code to calculate profit:
https://github.com/Varun487/BigData_Pair_Trading/blob/main/generate_pairs_orders_profit/calculate_profit.py

VISUALIZING A SAMPLE PAIR

For the pair **BAJAJ-AUTO and HEROMOTOCO**:

1. Close prices movement of pair of stock



In the figure above:

- **Blue line** - Symbol 1 Close price
- **Orange line** - Symbol 2 Close price

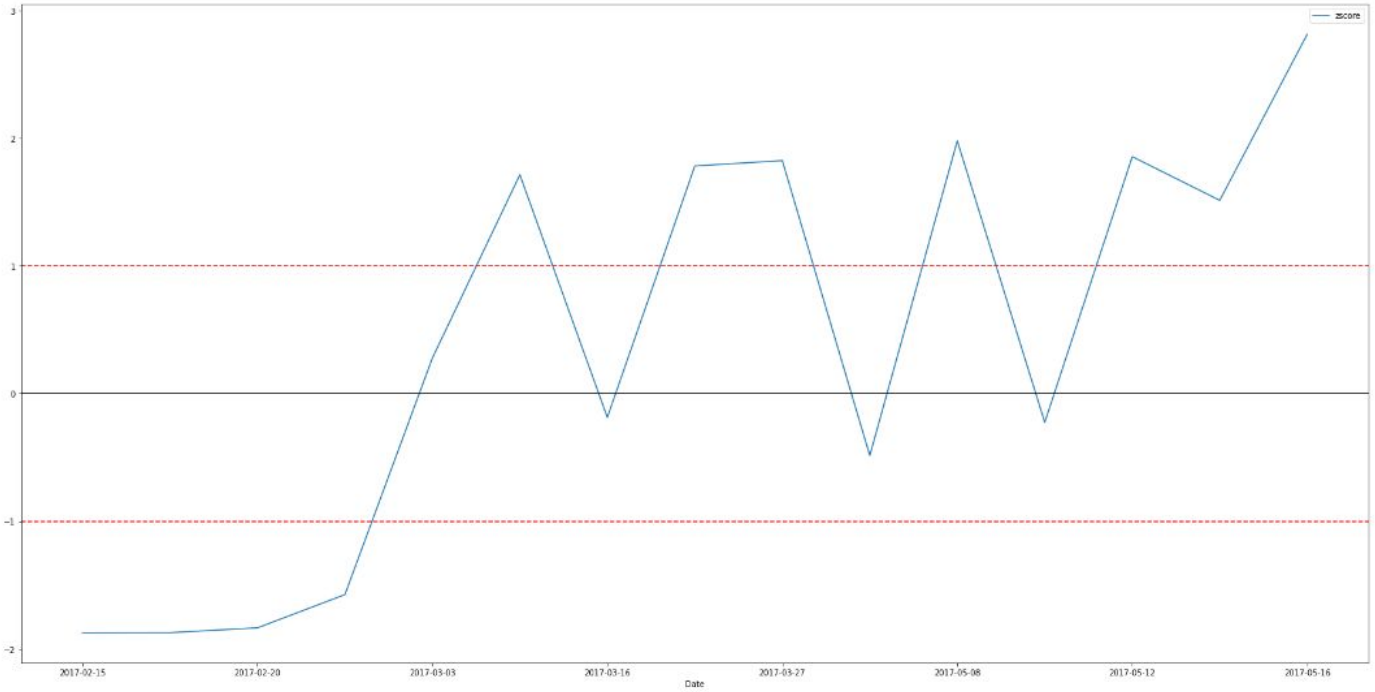
We can see from the figure above that the stocks move in tandem and visually form a good pair.

2. Calculated correlation matrix

	HEROMOTOCO_Close	BAJAJ-AUTO_Close
HEROMOTOCO_Close	1.000000	0.893185
BAJAJ-AUTO_Close	0.893185	1.000000

We can see from the correlation matrix above, that at a correlation value of 0.89, the pair is highly correlated.

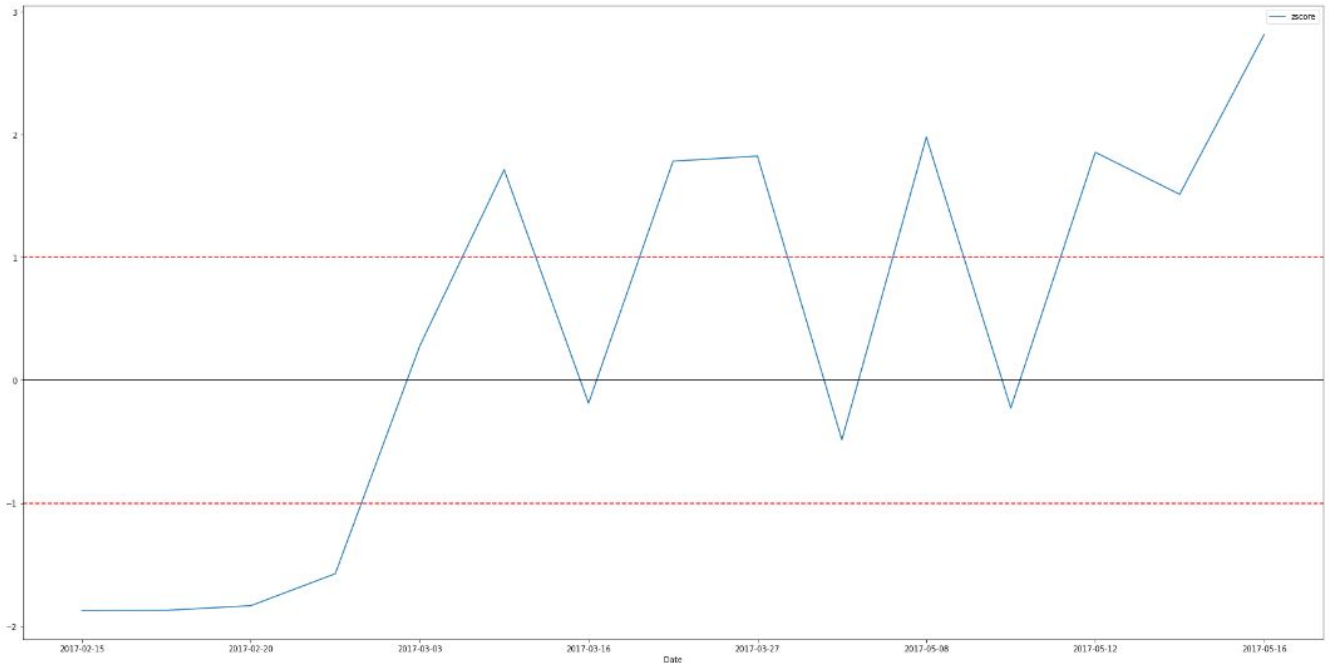
3. Visualisation of the Spread of the pair



In the graph above:

- **Blue line** is the zscore of the spread.
- **Black line** at 0 is mean.
- **Red dotted lines** represent 1 and 2 standard deviations above and below the mean respectively.

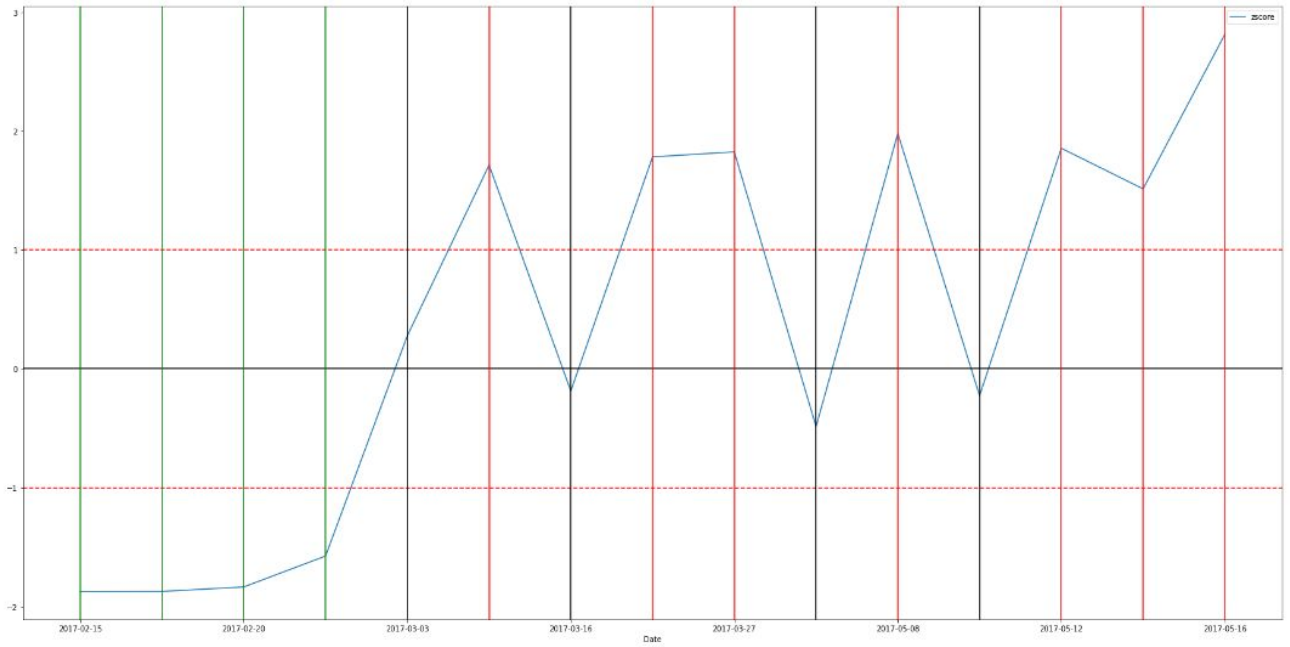
4.
Visualization of the Z-score of the Spread of the pair



In the graph above:

- **Blue line** is the zscore of the spread.
- **Black line** at 0 is mean.
- **Red dotted lines** represent 1 and 2 standard deviations above and below the mean respectively.

6. Visualization of Orders generated on the pair (on the z-score of the Spread of the pair)



In the figure above:

- **Blue line** - zscore of the Spread
- **Black horizontal line** at 0 - Mean
- **Red dotted horizontal lines** - at +1 and -1 standard deviations
- **Green vertical line** - represents long position taken on that day
- **Red vertical line** - represents short position taken on that day
- **Black vertical line** - represents getting out of all open positions till that point

7. visualizing profits/losses of orders generated

In the figure above:

- **Blue line** - Symbol 1 Profits
- **Orange line** - Symbol 2 Profits
- **Black line** at 0 - Profit Line
- All points below black line show a loss
- All points above black line show a profit

Conclusion

In this project, we have showcased how we can use **PYSPARK** to generate pairs and earn money in the real world. It shows the power of Big Data in evaluating large quantities of stock data to gain meaningful insights which can be exploited (Statistical Arbitrage) by companies and individuals to earn money in the stock market. In pairs trading, the ability of finding pairs with a good amount of correlation and cointegration can mean the difference between making a losing money. Thus, we have built an effective and profit making strategy using **PYSPARK** as a tool to gain an edge and beat the market.