# Pairs Trading

Motivation

# MSE criterion to form pairs

## Mean Squared Error

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

In simple words:
*'Error' here indicates deviation from some expected or predicted value, or from some model.*
*For example, if there's a bunch of points and you think "well, it's actually sort of like a straight line in shape, but there are small deviations from the line, it's noisy".*
*In order to find the 'best fitting' line, you'd say, "well, the line with which the deviations of each of the points is the least is the line". Fair enough?*

*If so, then you'd need to quantify the average ('mean') of the errors. But errors could be positive or negative, so you can't just sum them, as they can cancel each other. So you square them take the mean of that, and this gives you a nice number denoting how much deviation from your line is seen. Can be used to compare different lines or find the best one.*

# Reasoning behind employment of MSE to form pairs:

We use this approach because **it best approximates the description of how traders themselves choose pairs.** Interviews with pairs traders suggest that they try to find two stocks whose prices "**move together**". We then choose a matching partner for each stock by finding the security that minimizes the sum of squared deviations between the two normalized price series. Pairs are thus formed by exhaustive matching in normalized daily "price" space, where price includes reinvested dividends.
(Gatev, Goetzmann, & Rouwenhorst, 2006)

# Visualization and Interpretation of returns

Thresholds: 11 equidistant numbers from 0.0025 0.015

Entry Parameters: 21 equidistant numbers from 2 to 4

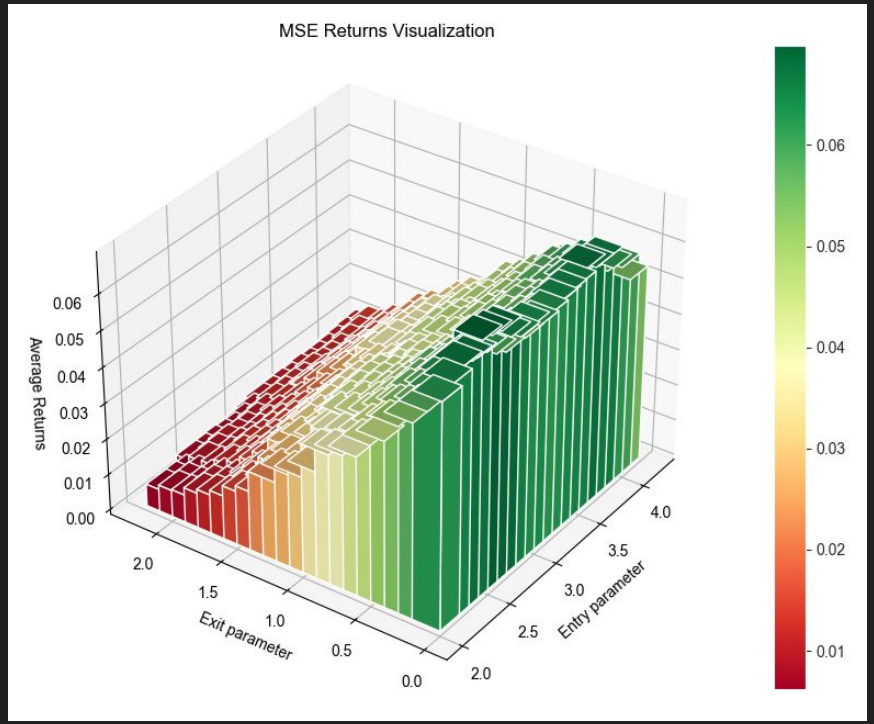
Exit Parameters: 21 equidistant numbers from 0 to 2

Entry and Exit parameters are varied and average returns across all 11 thresholds are computed. We obtain 441 returns corresponding to the 21 entry and 21 exit parameters. These 441 returns can be thought of as the height of the bars in the 3d bar plot.

Max returns = 6.978% (when entry = 2, exit = 0.4)

Min returns = 0.062% (when entry = 4 , exit = 0.3)

Standard deviation of the returns = 1.730%

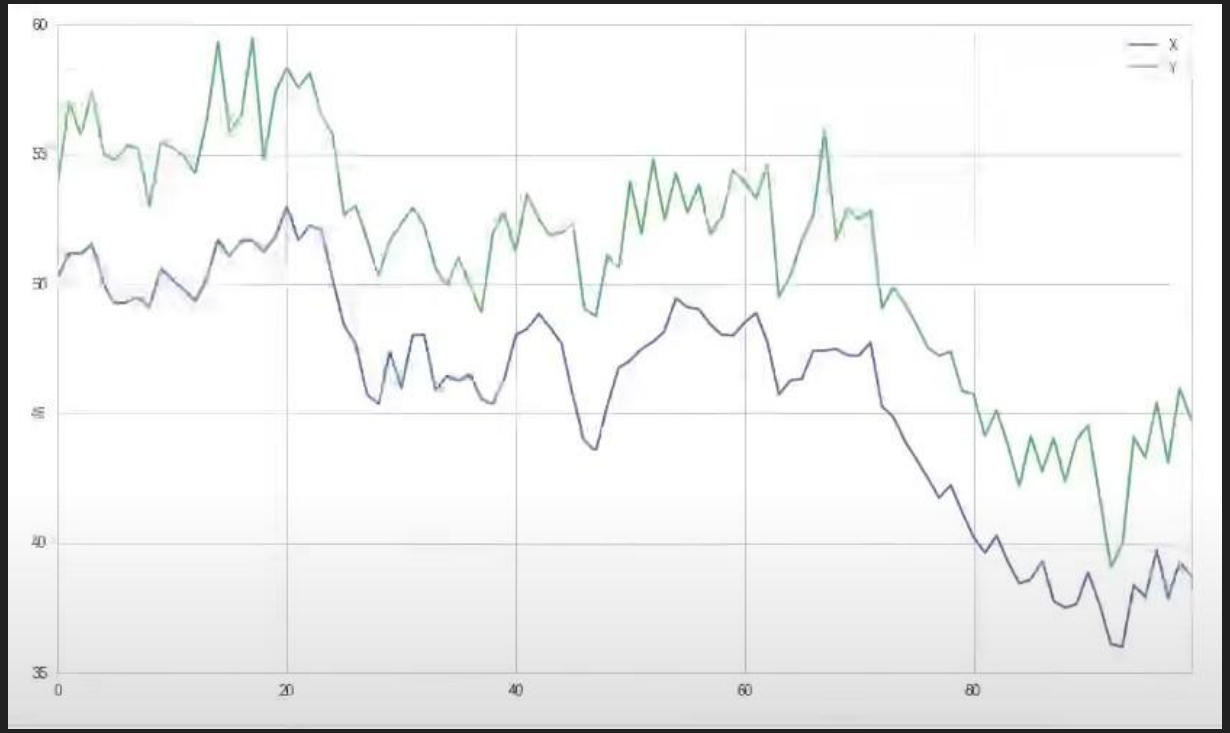


# Cointegration

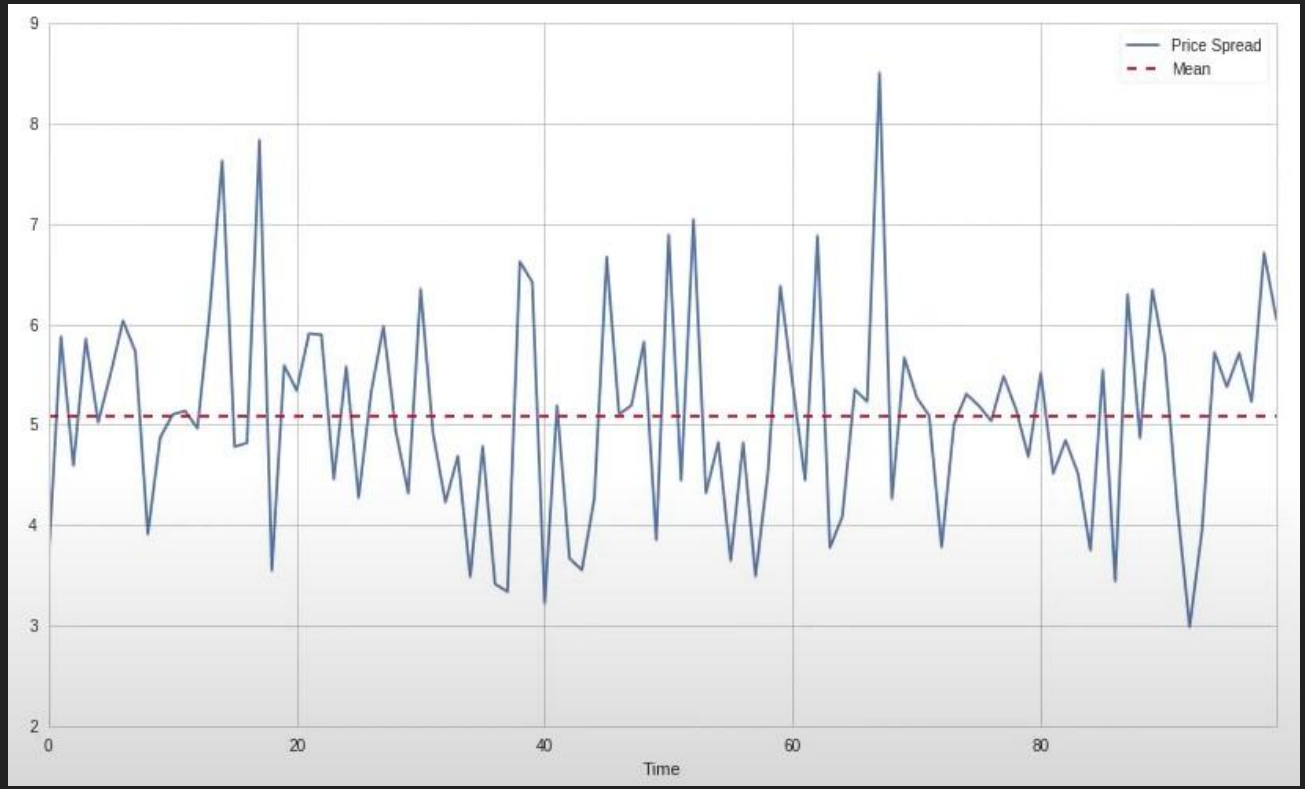
A cointegration test is used to establish if there is a correlation between several time series in the long term.

Cointegration tests identify scenarios where two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term. The tests are used to identify the degree of sensitivity of two variables to the same average price over a specified period of time

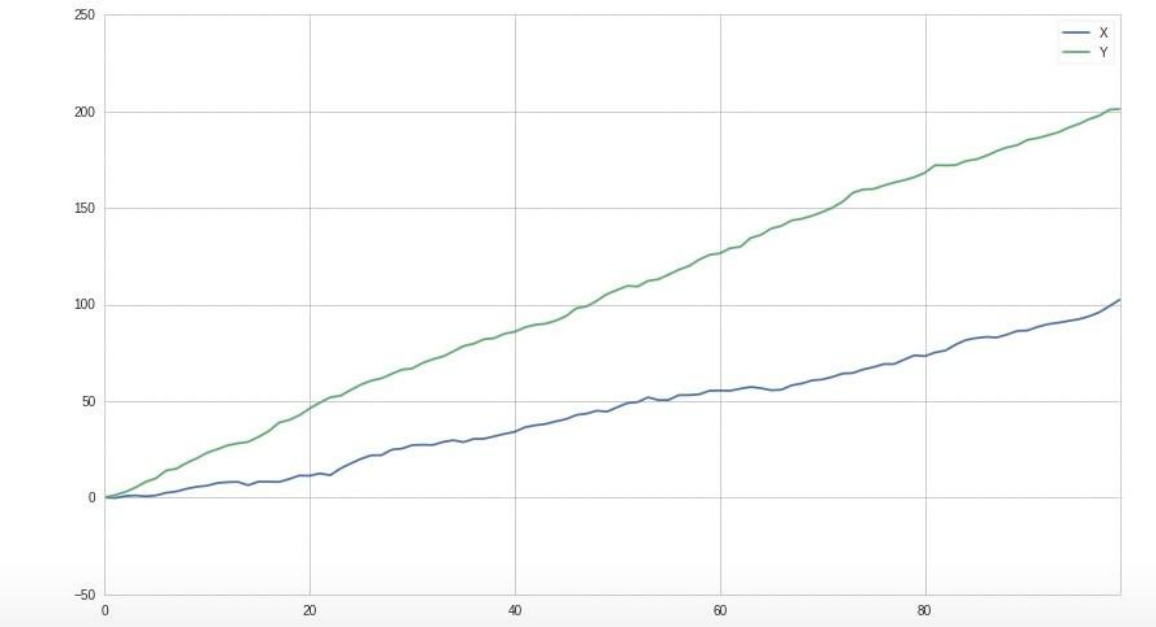
# “Distance” between two time series is nearly constant



# Mean reverting difference series



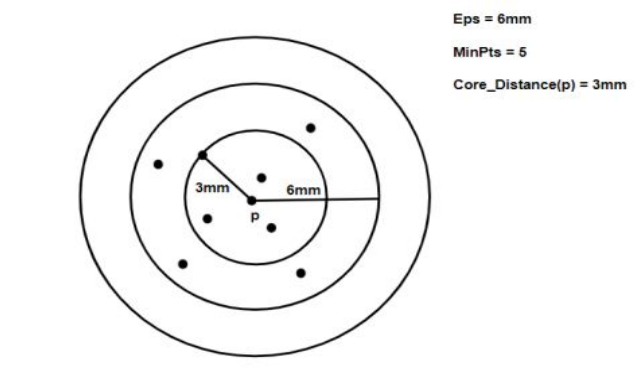
# Correlation ≠ Cointegration



# Engle-Granger test

The Engle-Granger Two-Step method starts by creating residuals based on the static regression and then testing the residuals for the presence of unit roots. It uses the Augmented Dickey-Fuller Test (ADF) or other tests to test for stationarity units in time series. If the time series is cointegrated, the Engle-Granger method will show the stationarity of the residuals.

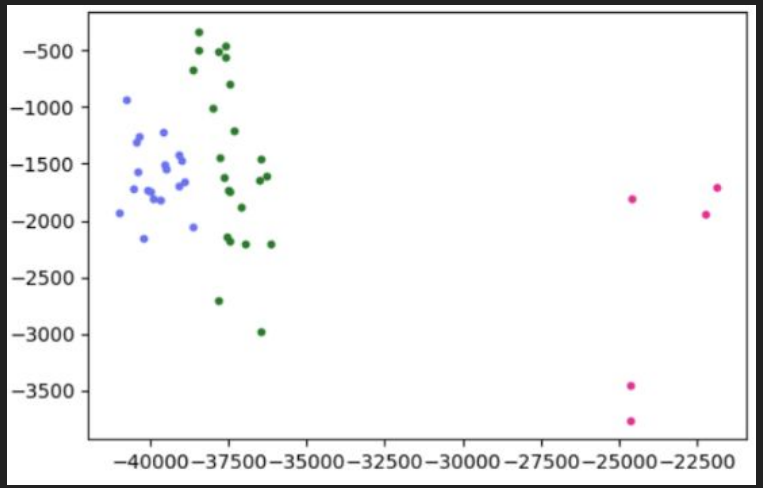
# OPTICS Clustering Algorithm




Core Point is when there are more than minPts present in the neighborhood(epsilon-radius circle) of a point.

Reachability distance is calculated between the other points and a core point and then assigned a cluster.

Methodology followed:
1) Calculate the daily returns of the stocks
2) Perform PCA to reduce the dimensionality
3) Perform OPTICS clustering on the PCA reduced data
4) Use cointegration with a cluster to rank the pairs



# References

1. Pairs Trading: Performance of a Relative Value Arbitrage Rule