DOCUMENTATION: QUANT Problem Statement

A. PAIR IDENTIFICATION AND SELECTION

A pair is selected based on their good statistical arbitrage opportunities over time. Correlation and cointegration are 2 important terms which we calculated to keep in check if the stocks follow the required relative price movements, i.e. in line with good statistical arbitrage.

Correlation: Correlation describes the relation between variables ad is quantified by the correlation coefficient ρ, ranging from -1 to +1. The value of +1 indicates a perfect positive correlation between the two variables, -1 indicates a perfect negative correlation and 0 means there is no correlation.

$$\text{Correlation}(X,Y) = \rho = \text{COV}(X,Y) / \sigma(X)\sigma(Y)$$

Where,
COV is covariance
σ is standard deviation

Cointegration: Cointegration is a statistical property of two or more time-series variables which indicates if a linear combination of the variables is stationary.Parameters such as mean and variance do not change over time.

$$\text{Spread} = Y - n*X$$

Where,
n is the hedge ratio (It is ideal when spread= 0)

Libraries imported for downloading data for the past years is **yfinance.** This library provides extensive data regarding the stocks' opening price, closing price, adjacent close price, low price and high price over the years for different stocks/tickers.

ANALYSIS TOOLS FOR PAIR SELECTION:

Data from sites NSE(Nifty 50) are taken.Pairs of 2 stocks are checked for their sharpe ratio and CAGR. The pair with highest sharpe ratio and CAGR is selected. Following that, we proceed further in the trading strategy.

If the co-integration test meets our threshold statistical significance (in our case 5%), then that pair of tickers will be stored in a list for later retrieval.

B.TRADING STRATEGY AND SIGNAL GENERATION METHOD

Statistical arbitrage works to see when the trade should be done. The spread of both the stocks is found and then it is seen where a significant deviation from the mean is observed. This will help us to categorize among the correlated assets, one will be the "lead" asset and the other would be the "lag" asset. The lead asset typically outperforms the lag asset.

The assumption behind this strategy is that the spread from pairs that show properties of co-integration is mean reverting in nature and therefore will provide arbitrage opportunities if the spread deviates significantly from the mean.

The lead asset would grab a long position(buy the stock, in hope that its price would increase later) and the lag asset would grab the short position(sell the stock, hoping that the price of the stock would fall) and this difference would provide us with the profit or loss depending on how good our trade is.

Backtesting the pairs trading strategy by calculating the cumulative returns, Sharpe ratio, and the **Compound Annual Growth Rate(CAGR)** for a pair of assets based on historical price data provided in a DataFrame.

We ran a regression analysis to find the hedge ratio (hr) between the two assets using **Kalman Filter**.

Kalman filter for dynamic hedge ratio calculation

Kalman filtering, also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies.

It produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each time-frame.

Then, calculated the spread series as **y - (x * hr)**.

This is useful for computing the moving average if that's what we are interested in, or for smoothing out estimates of other quantities.

Z-score is a measure of how many standard deviations the current spread is from its historical mean.
We then defined entry and exit Z-score thresholds for long and short positions. Identified long entry and exit points based on the Z-score criteria,short entry and exit points based on the opposite Z-score criteria.

C. TRADING SIGNALS GENERATED AND POSITION SIZING

Trading logic

1. Calculate the spread of each pair (Spread = Y – hedge ratio * X)
2. Using Kalman Filter Regression Function to calculate hedge ratio
3. Calculate z-score of ‘s’, using rolling mean and standard deviation for the time period of ‘half-life’ intervals. Save this as z-score
4. Using half-life function to calculate the half-life
5. Define upper entry Z-score = 2.0, lower entry Z-score = 2.0, exit Z-score = 0.0
6. When Z-score crosses upper entry Z-score, go SHORT; close the position with Z-score return exit Z-score
7. When Z-score crosses lower entry Z-score, go LONG; close the position with Z-score return exit Z-score
8. Back-test each pair, and calculate the performance statistics, each as max drowns down Sharpe ratio

9. Build up portfolios with equal market value distribution, each pair has the same market value

Backtesting

The back-test engine follows the steps:

1. Calculate Spread = Y – hedge ratio * X
2. Using Kalman Filter Regression Function to calculate hedge ratio
3. Calculate z-score of ‘s’, using rolling mean and standard deviation for the time period of ‘half-life’ intervals. Save this as z-score
4. Using half-life function to calculate half life
5. Define upper entry Z-score = 1.25, lower entry Z-score = -1.25, exit Z-score = -0.5
6. When Z-score crosses upper entry Z-score, go SHORT; close the position with Z-score return exit Z-score
7. When Z-score crosses lower entry Z-score, go LONG; close the position with Z-score return exit Z-score

D. PORTFOLIO PnL

An initial capital of $100,000 is taken and number of shares to be bought for each stock are calculated based on initial capital and the prices of the stocks at the beginning of the trading period.

PnL is then calculation for Stock 1:
DataFrame called portfolio is created to track various aspects of the PnL for stock 1.

The holdings for stock 1 are tracked based on the cumulative positions (signals) and stock prices.The remaining cash after buying and selling stock 1 is tracked. The total value for stock 1, which is the sum of holdings and cash and the daily returns for stock 1 is calculated.

PnL Calculation for Stock 2:
Similar to stock 1, all the same factors are also tracked and calculated for stock 2 as well.

The total PnL is calculated by adding the PnL of stock 1 and stock 2, which is already stored.
DataFrame is cleaned by removing the NaN values and then returned back.

E. PERFORMANCE METRICS

Sharpe Ratio:

It help to assess the risk-adjusted return of an investment or portfolio.It is often used to compare different investments or portfolios and assess which one provides the best risk-adjusted return.

Sharpe Ratio (SR) = (Rp - Rf) / σp

Where,

Rp: The average return of the investment or portfolio.
Rf: The risk-free rate of return.
σp: The standard deviation of the investment's or portfolio's returns.

A higher sharpe ratio is generally preferred, as it indicates a better risk-adjusted return.

CAGR:

It is used to measure the annual growth rate of an investment or asset over a specific period, assuming that the investment has been compounding.It gives you a standardized way to evaluate the return on an investment or asset, even when the growth or returns are not uniform over time

CAGR = (Ending Value / Beginning Value) ^ (1 / n) - 1

Where,

Ending Value= value of the asset at the end of the specified time period.
Beginning Value= initial value of the asset at the beginning of the specified time period.
n= The number of years in the time period.

It is generally used to assess the historical performance of investments or to project future returns.