Pair Trading

This is a very common starting strategy for a lot of traders, which uses fundamental concepts like mean reversion and statistical arbitrage.

Given the list of stocks from the US IT sector in the NYSE Stock Exchange:

- Tesla (TSLA)
- Amazon (AMZN)
- Apple (AAPL)
- Alphabet Inc (GOOGL)
- Microsoft (MSFT)

Perform the following tasks:

- **Selection and Rationale:** Choose any two stocks from the above list and explain your selection criteria, including sectoral trends and any fundamental differences between the two companies. Include any correlation analysis which you may have done.
- **Data Collection and Preprocessing:** Fetch historical price data for the chosen stocks (01/01/2019 – 31/12/2023) using the `yfinance` library (Use the stocks from the NYSE stock exchange). Perform data preprocessing such as handling missing values, normalizing prices, and visualizing price movements.
- **Backtesting the Strategy:** Develop and backtest a pair trading strategy for the chosen stocks using the spread of their prices. Calculate key performance metrics such as returns, Sharpe ratio, and maximum drawdown.
- **Performance Analysis:** Validate the robustness of your strategy and discuss potential improvements.

The Pair Trading question is answered in the form of a report below.

Pair trading question's answer explained step by step:

Note: The complete Python code, including data preprocessing and backtesting functions, is attached at the end for reference.

- **Selection of the stocks:**

⇒ I considered 10 stocks from the US IT sector and proceeded to figure out how to choose the required pair of stocks. I chose GOOGL and ACN as my pair due to its high correlation and low cointegration values.

⇒ Correlation measures how similarly two stock prices move—values close to +1 indicate they move together, while values near −1 suggest they move in opposite directions.

⇒ Cointegration means that while two stock prices may individually wander, their difference stays stable over time—suggesting a long-term equilibrium relationship. If the p-value was less than 0.05, we conclude that cointegration exists between the two sets of data/time series.

⇒ The image below shows all values of the above for all combinations of those 10 companies.

```
--- Pairs sorted by Correlation (highest to lowest) ---
AAPL - MSFT | Correlation: 0.9717 | Cointegration p-value: 0.3672
GOOGL - ACN | Correlation: 0.9677 | Cointegration p-value: 0.0021
MSFT - ACN | Correlation: 0.9494 | Cointegration p-value: 0.1843
GOOGL - MSFT | Correlation: 0.9418 | Cointegration p-value: 0.4897
AAPL - ACN | Correlation: 0.9318 | Cointegration p-value: 0.2295
TSLA - ACN | Correlation: 0.9225 | Cointegration p-value: 0.1577
TSLA - GOOGL | Correlation: 0.9137 | Cointegration p-value: 0.1312
MSFT - ORCL | Correlation: 0.9098 | Cointegration p-value: 0.2166
MSFT - NOW | Correlation: 0.9067 | Cointegration p-value: 0.4161
GOOGL - NOW | Correlation: 0.9038 | Cointegration p-value: 0.2889
AAPL - GOOGL | Correlation: 0.9031 | Cointegration p-value: 0.5168
TSLA - AAPL | Correlation: 0.8982 | Cointegration p-value: 0.4837
ACN - NOW | Correlation: 0.8965 | Cointegration p-value: 0.2863
AMZN - CRM | Correlation: 0.8763 | Cointegration p-value: 0.0661
TSLA - MSFT | Correlation: 0.8751 | Cointegration p-value: 0.6482
TSLA - NOW | Correlation: 0.8729 | Cointegration p-value: 0.2222
AAPL - ORCL | Correlation: 0.8716 | Cointegration p-value: 0.3723
AAPL - NOW | Correlation: 0.8636 | Cointegration p-value: 0.4946
CRM - NOW | Correlation: 0.8416 | Cointegration p-value: 0.1145
ORCL - ACN | Correlation: 0.8398 | Cointegration p-value: 0.4603
GOOGL - ORCL | Correlation: 0.8317 | Cointegration p-value: 0.4098
AMZN - NOW | Correlation: 0.8276 | Cointegration p-value: 0.4599
ORCL - IBM | Correlation: 0.8207 | Cointegration p-value: 0.1982
ORCL - NOW | Correlation: 0.7469 | Cointegration p-value: 0.5223
MSFT - IBM | Correlation: 0.7213 | Cointegration p-value: 0.4327
AAPL - IBM | Correlation: 0.7158 | Cointegration p-value: 0.4083
TSLA - AMZN | Correlation: 0.6978 | Cointegration p-value: 0.6348
TSLA - ORCL | Correlation: 0.6975 | Cointegration p-value: 0.6347
IBM - ACN | Correlation: 0.6890 | Cointegration p-value: 0.5429
GOOGL - CRM | Correlation: 0.6728 | Cointegration p-value: 0.7400
AMZN - GOOGL | Correlation: 0.6716 | Cointegration p-value: 0.6330
GOOGL - IBM | Correlation: 0.6451 | Cointegration p-value: 0.5960
CRM - ACN | Correlation: 0.6314 | Cointegration p-value: 0.5100
AMZN - ACN | Correlation: 0.6314 | Cointegration p-value: 0.6150
TSLA - CRM | Correlation: 0.6257 | Cointegration p-value: 0.6750
MSFT - CRM | Correlation: 0.6201 | Cointegration p-value: 0.8477
AMZN - MSFT | Correlation: 0.5967 | Cointegration p-value: 0.7081
AMZN - AAPL | Correlation: 0.5553 | Cointegration p-value: 0.6670
AAPL - CRM | Correlation: 0.5396 | Cointegration p-value: 0.8102
IBM - NOW | Correlation: 0.5042 | Cointegration p-value: 0.6211
TSLA - IBM | Correlation: 0.5041 | Cointegration p-value: 0.5485
ORCL - CRM | Correlation: 0.4648 | Cointegration p-value: 0.8381
AMZN - ORCL | Correlation: 0.3532 | Cointegration p-value: 0.6084
CRM - IBM | Correlation: 0.1427 | Cointegration p-value: 0.6162
AMZN - IBM | Correlation: 0.0794 | Cointegration p-value: 0.5880
```

Figure 1: All values of coint and correl from all combinations of companies chosen

⇒ From here, I chose the pair with the higher correlation and lower co-integration p-value. That seemed to be best suitable for pair trading.

⇒ The reason I did not choose AAPL and MSFT as the pair is that even though their correlation value was the highest, their cointegration value was not low enough.

⇒ Once normalized, the graphs of both GOOGL and ACN also semed to move similar from the date 01/01/2019 to 31/12/2023. If they are not normalized then we would not be able to compare conveniently.

⇒ The image of the graphs of both GOOGL and ACN after normalizing with respect to one is given below.

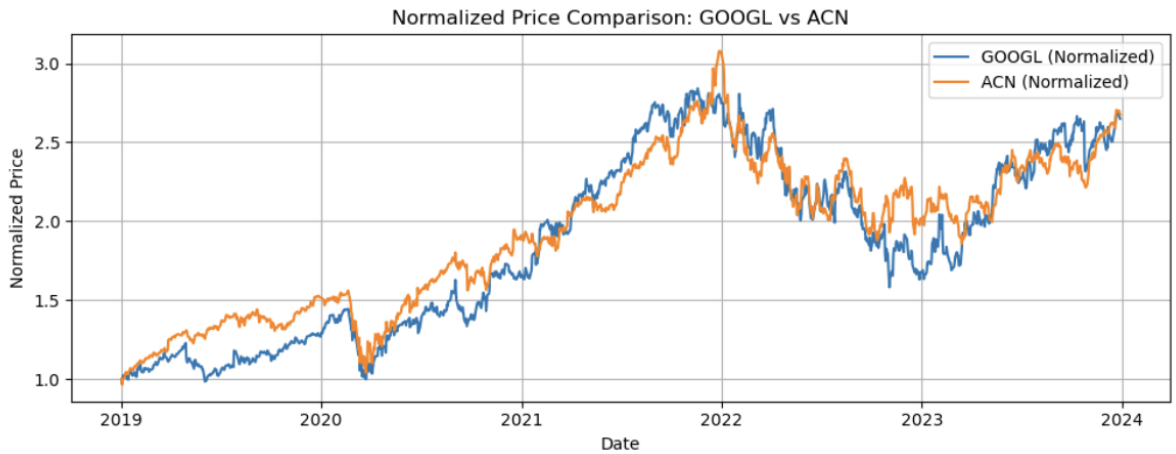

Figure 2: GOOGL and ACN price lines after normalizing

⇒ As we can see from the graph, both the price lines are very close to eachother and sometimes go apart, but come back towards eachother at some point. This shows that it is an essential pair for pair trading.

- **Data Collection:**

⇒ I did fetch the historical price data for the range of given dates from the yfinance library as can be seen from my code.

- **Data Preprocessing:**

⇒ Handling missing values (NaN) was done by doing "data.dropna(inplace=True): this in python.

⇒ Normalizing prices was done by dividing the values across all days by the value of day 1.

⇒ Visualizing price movements was done by plotting graphs primarily using matplotlib library in python.

- **Spread:**

⇒ Spread captures the difference in behaviour of the two stocks.

⇒ I first made a simple spread on the normalised prices by using the simple difference of both the stocks(I subtracted ACN normalized prices from GOOGL normalized prices).

⇒ For a more accurate spread, I introduced beta-which was found using linear regression by using GOOGL as the dependant variable and ACN as the independant variable. "How does GOOGL's price behave as ACN's price changes"

⇒ Beta is the slope from a linear regression between the two prices

⇒ The beta value was found to be 1.1476.

⇒ I have shown the simple spread plot and the beta spread plot below.

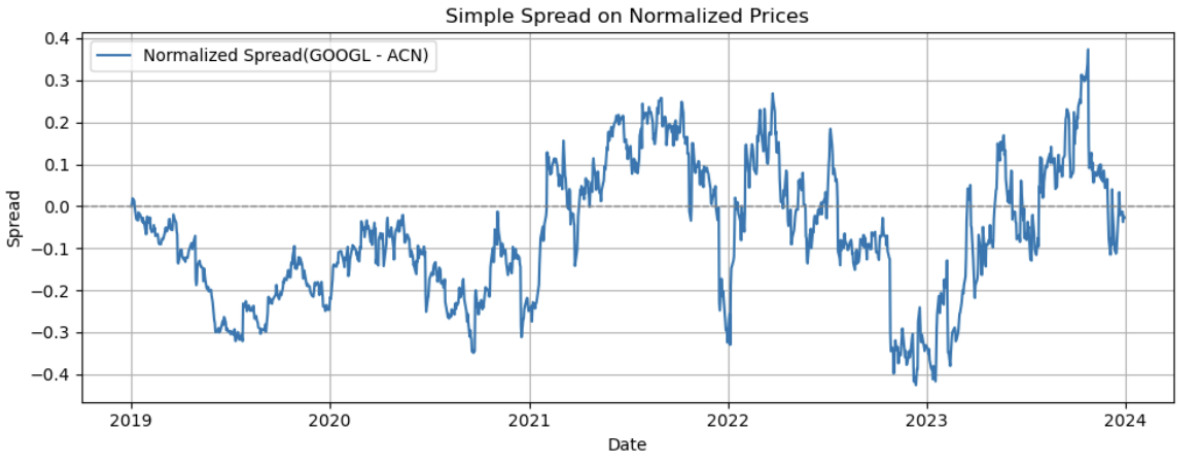

Figure 3: GOOGL and ACN simple spread

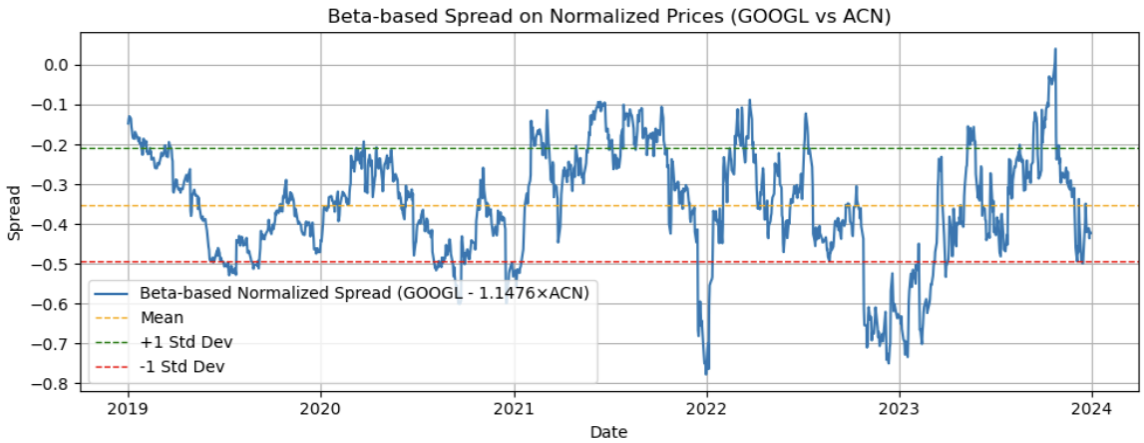

Figure 4: GOOGL and ACN beta spread with mean and +/-1 standard deviations shown

- **Strategy:**

⇒ Essentially what we do is, if the spread of the data is more than +1 std dev, then we short GOOGL and long ACN as we are using GOOGL- beta* ACN as the formula for spread. Here I also introduce the concept of signals.

- When the data is more than +1 std dev, that means GOOGL price has risen and will most likely come back down soon. So we short GOOGL when the price rises and the price line graph of GOOGL and ACN diverges. Shorting GOOGL

means we will sell GOOGL stocks now and buy them later on when the price falls back down, which means we will profit from this. We long ACN by buying now and selling later in hopes that in case ACN's price increases, it will come back up in some time, again being profitable.
I give signal value as +1 if the data is more than +1 std dev, indicating that performing short for GOOGL and long for ACN will be profitable at that point in time.

- When the data is less than -1 std dev, that means ACN price has risen and/or GOOGL's price has lowered and will most likely come back to normal soon. So we short ACN when the price rises and the price line graph of GOOGL and ACN diverges. Shorting ACN means we will sell ACN stocks now and buy them later on when the price falls back down, which means we will profit from this. We long GOOGL by buying now and selling later in hopes that in case GOOGL's price increases, it will come back up in some time, again being profitable.
I give signal value as -1 if the data is less than -1 std dev, indicating that performing long for GOOGL and short for ACN will be profitable at that point in time. I have given the graph with the signal values indicating when to do what kind of pair trading below:

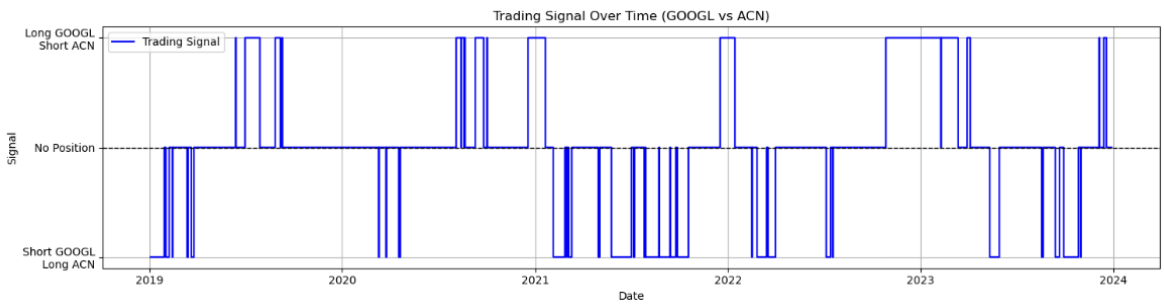

Figure 5: GOOGL and ACN beta spread with mean and +/-1 standard deviations shown

⇒ What do I do with the spread? If the spread is more than +1 std dev(signal = +1) or less than -1 std dev(signal = -1) that means the stock prices of GOOGL and ACN have diverged enough in order to get substantial returns from pair trading-i.e, it is not the small trend variation that could last permanently- instead, it is an abnormal variation which is very likely to come back to its usual trends. This means it is a good opportunity for pair trading.

- **Backtesting the Strategy and Performance Analysis:**

⇒ I ran a few tests and got the values of a few performance metrics to see how well my model is performing. I then compared it to the S&P 500 graph to check if it performs better or worse.

⇒ Cumulative Return: This is the total percentage gain or loss of an investment over a period, showing how much it has grown relative to the starting value. I am doing all the calculations by taking 1 rupee as the sample amount.
I have calculated:
Total Return Multiplier: $\frac{\text{Final}}{\text{Initial}} = 3.22\text{x}$
Total Return Percentage: 221.78%
Analysis:A cumulative return of over 200% indicates strong performance. Since the return multiplier is 3.22x, the strategy significantly outperformed typical market benchmarks like the S&P 500 over the same period.(will elaborate more on this in a later section) The graph is given below.

Figure 6: GOOGL and ACN beta spread with mean and +/-1 standard deviations shown

⇒ Sharpe Ratio : This tells you how much extra return you're getting for the risk you're taking with an investment.
Sharpe Ratio=1.7559
Analysis:A Sharpe Ratio of 1.7559 indicates a strong risk-adjusted performance of the pair trading strategy. Typically, a Sharpe Ratio above 1.0 is considered good, suggesting that the strategy is generating returns in excess of the risk-free rate with acceptable levels of risk. A value above 1.5 is viewed as very good, and anything approaching or exceeding 2.0 is considered excellent. In this case, the ratio of 1.7559 demonstrates that the trading strategy between GOOGL and ACN delivers consistent and substantial returns for the amount of volatility involved.

⇒ Maximum Drawdown: It is the maximum observed loss from a peak to a trough of an investment before a new peak is reached.
Max Drawdown= -9.54% (negative indicates the decrease))
Analysis:A max drawdown below 10% (modulus) is relatively low, which is favorable as it suggests the strategy doesn't experience large losses. A value of -9.54% shows the downside risk was well-controlled.

⇒ Volatility (Annualized) : It is a measure how much the investment's value goes up and down in a year, i.e, how much it fluctuates — higher volatility means more risk.
Volatility (Annualized)=0.1389
Analysis: Annualized volatility between 10% to 20% is typical for moderately risky strategies. A value of 13.89% suggests the strategy has a balanced risk profile — not too volatile, yet active enough to generate returns.

⇒ Alpha (Annualized) : It is the extra return your strategy makes compared to the market, after adjusting for risk.
Alpha (Annualized)= 0.2459
Analysis:A positive alpha indicates outperformance over the market. An alpha of 0.2459 (or 24.59%) annually is excellent and shows that the strategy is generating strong excess returns over the benchmark.

⇒ Beta vs S&P 500 : It shows how closely your investment moves with the S&P 500 — a beta of 1 means it moves the same way.
Beta vs S&P 500=0.0123
Analysis:A beta close to 0 implies very low correlation with the market, which means the strategy behaves independently of overall market trends. A beta of 0.0123 suggests great diversification benefits.

⇒ Profit Factor: The total money made divided by the total money lost — higher means more profitable trades.
Profit Factor=1.7675
Analysis:A profit factor above 1.5 is considered good, showing that gains significantly outweigh losses.

⇒ I am now comparing how much my investment of 1 rupee has grown over time versus how much the SP 500 has grown over the same period assuming 1 rupee again. The following is the graph:

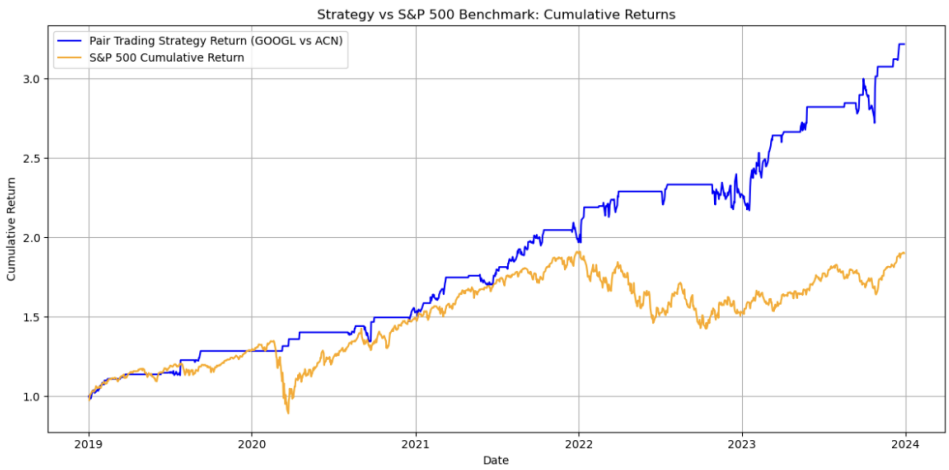

Figure 7: My pair trading strategy growth compared to S&P 500 growth

⇒ As is clearly visible from the graph, 3.22x is the cumulative return I have from my strategy, whereas the cumulative return from the S&P 500 are only 1.90x over the same period, demonstrating a significant outperformance done by the pair trading approach.

- Scope for Improvement

⇒ Z-score usage : Using z-score to compare the stock prices

⇒ ADF test : Even if two stocks are cointegrated, it's useful to test whether the spread between them actually moves around a fixed average (stationary), using tests like these.

⇒ Use a rolling window (e.g., 30 or 60 days) to calculate mean and standard deviation of the spread can help improve results.

Code Attachment

You can download the code from the following GitHub repository:
`Download fin_club_code.ipynb`

Additional Information: Instances when Pair trading could have worked for GOOGL-ACN in the past

1. March 2020 - COVID-19 Market Crash

- **Market Context:** The onset of the COVID-19 pandemic triggered a rapid and broad selloff across global equities.
- **Observed Behavior:** ACN's stock price fell more sharply than GOOGL, resulting in the spread increasing significantly and crossing the threshold standard deviation from the mean.
- **Trading Signal:** *Short GOOGL, Long ACN.*
- **Outcome:** As markets rebounded, particularly in the technology sector, ACN recovered quickly, leading to a narrowing of the spread and a profitable mean-reversion.

2. November 2021 - Tech Sector Peak

- **Market Context:** Late 2021 witnessed record highs for many tech stocks, driven by post-COVID earnings strength and low interest rates.
- **Observed Behavior:** GOOGL surged significantly, while ACN lagged, causing the spread to exceed the threshold standard deviation.

- **Trading Signal:** *Short GOOGL, Long ACN.*
- **Outcome:** A correction in early 2022 hit GOOGL harder than ACN, leading the spread to revert to its mean, thus closing the trade profitably.