Identifying Strong Pairs Trading Strategies

Ryan Lagasse, Sahil Gandhi

Objective Statement:
The goal of this project is to develop a pairs trading strategy by identifying pairs of stocks or other financial instruments that exhibit a strong historical correlation in their price movements. The strategy will aim to capitalize on temporary deviations from the established relationship between the pair, allowing for potential profit opportunities when the prices converge back to their typical pattern.

Data Collection:
I will obtain historical price data for a set of stocks or other financial instruments from a reliable source, such as Yahoo Finance or a financial data provider. The data will cover a sufficiently long time period to capture the historical relationships between the asset pairs.

Data Cleaning:
Using Python's pandas library, I will handle any missing or erroneous data in the dataset. This may involve filling in missing values, removing outliers, or addressing any other data quality issues to ensure the dataset is clean and ready for analysis.

Data Visualization:
I will create visualizations, such as scatter plots and time series plots, to explore the relationships between the asset pairs. These visualizations will help identify pairs that exhibit strong historical correlations and may be suitable candidates for pairs trading strategies.

Data Modeling:
To identify pairs with strong correlations, I will calculate various statistical measures, such as the correlation coefficient, cointegration tests, and distance metrics. These measures will help quantify the strength of the relationship between the asset pairs and identify pairs that have maintained a consistent pattern over time. I will explore the use of machine learning algorithms to identify pairs with strong historical correlations. Some potential approaches include:

- Clustering algorithms: Techniques like k-means clustering or hierarchical clustering can be used to group stocks or financial instruments based on their price movement patterns. Pairs within the same cluster are likely to exhibit strong correlations and could be candidates for pairs trading strategies.
- Dimensionality reduction: Methods like Principal Component Analysis (PCA) or t-SNE can be used to reduce the dimensionality of the data and identify the most

significant components or features that drive the relationships between the asset pairs.
- Supervised learning: If historical data on successful pairs trading strategies is available, supervised learning algorithms like random forests or support vector machines could be trained to identify patterns and characteristics of profitable pairs.

Analysis:
I will develop a pairs trading strategy based on the identified pairs with strong historical correlations. This may involve setting thresholds for when to enter and exit trades based on the deviation from the established relationship. I will backtest the strategy using the historical data to evaluate its performance and potential profitability. Once potential pairs have been identified using statistical measures and machine learning techniques, I will develop a pairs trading strategy based on the identified pairs with strong historical correlations. This may involve:

- Training machine learning models: I will train machine learning models, such as decision trees or neural networks, to predict when the prices of the asset pairs are likely to converge or diverge from their established relationship. These models can be trained on historical data, incorporating features like price differences, momentum indicators, or other technical analysis metrics.
- Entry and exit signals: Using the trained machine learning models, I will define rules or thresholds for when to enter and exit trades based on the predicted price movements of the asset pairs.
- Backtesting and evaluation: I will backtest the pairs trading strategy using the historical data to evaluate its performance and potential profitability. This will involve simulating trades based on the entry and exit signals generated by the machine learning models and calculating metrics like cumulative returns, Sharpe ratio, and maximum drawdown.

Throughout the project, I will document my process, including any challenges faced, decisions made, and insights gained. I will provide a detailed write-up explaining the steps taken, the rationale behind the choices made, and an analysis of the strategy's performance and potential limitations.

Submission:
I will create a public GitHub repository to host my code, data, and documentation. The repository will be organized and well-structured, with clear README files and comments within the code to facilitate understanding and reproducibility.