# Do Corporate Insiders Know Something We Don't?

*Thomas MacPherson, Kirtland Corregan, Rami Abushamalah*

**Overview:** When corporate executives, directors, or owners legally buy shares of their own company, it is known as insider trading (Hussain, 2022). This study examines whether open-market insider purchases, reported on SEC Form 4, can reliably predict six-month outperformance relative to the broader market. By focusing solely on buy-side transactions, **we treat insider purchases as signals of confidence in the firm's health and growth prospects**. We hypothesize that the majority of insider activity reflects random noise and yields returns similar to the broader market. However, we also hypothesis that certain insiders consistently achieve significant excess returns (≥20%). After data aggregation and cleaning, we compared post-purchase stock returns and found that most insider trades underperform the market benchmark, while a subset of individuals showed repeatable excess returns. Limitations in the analysis include survivorship bias, as failed companies that are delisted from the exchange may be excluded from the dataset.

## Motivation

**Problem:** Before including insider activity in active investment management, we aim to assess whether insider buying activity can be systematically leveraged to generate excess returns over the broader market. This evaluation will inform us on whether these transactions contain actionable signals that justify further model development.

**Approach:** We use historical Form 4 filings from 2006 through Q1 2025. For each open-market insider purchase, we calculate the stock's return over the subsequent six months and compare it to the return of the SPY ETF over the same period to find excess returns

## Data Sources

**Source 1:** Form 4 filings, available from the SEC website, provide essential details related to insider purchases and sales but do not include stock price performance around the transaction date. To test our hypothesis, we integrate this data with external market data from public sources.

**Source 2:** We use the open-source Yahoo! Finance API (yfinance) to retrieve stock price data and company information surrounding each insider transaction. We also include the SPY ETF price data as a benchmark for the S&P 500 broader market performance.

## Data Manipulation

**Manipulation:** Complex datasets pose several unique challenges. Insider roles and title formats vary widely, transaction and report dates often do not align, and stock prices may contain errors. We applied standard data science techniques such as regular expressions, grouping, filtering, formatting, removing, and replacing to clean and standardize our dataset.

**Exploration:** Assessing the quality and accuracy of our information was vital to draw conclusions. Our goal was to derive and present one meaningful insight from our large dataset.

## Analysis

**Positions:** Our analysis confirmed that most insider trading activity resembles random noise. In fact, the median return following insider purchases underperformed the broader market. No insider position, including high-ranking financial roles like Chief Financial Officer, consistently outperformed others.

**Individuals:** However, a small subset of individual insiders generated significant alpha. We identified 388 "elite alpha insiders", each with three or more purchases that produced raw six-month excess returns of 20% or more.

# Motivation

## Background

The goal of active investment management is to generate returns that exceed a chosen market benchmark on a risk-adjusted basis, commonly referred to as “generating alpha” **(Fig. 1)**. Investors often measure success by comparing their returns against a broad market index like the S&P 500, which tracks a diversified portfolio of 500 large-cap U.S. companies.

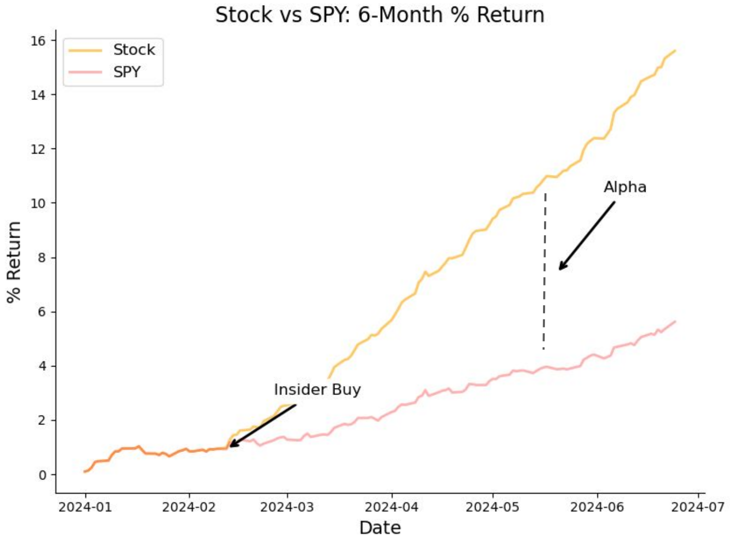


**Figure 1 -** Hypothetical example of alpha

Insider trading, in the legal sense, refers to open-market purchases or sales of a company’s stock by individuals who are likely to have privileged insight into the company’s future performance. These insiders include but are not limited to, CEOs, CFOs, board members, and major shareholders. While trading based on confidential, nonpublic information is illegal, the U.S. Securities and Exchange Commission (SEC) requires all legal insider trades to be disclosed publicly within two business days via Form 4 filings.

**“Can Insider activity help predict which stocks will outperform the broader market?”**

**This study focuses solely on insider purchases, which are more likely to reflect confidence in the company’s future.** By contrast, insider sales may be driven by personal reasons unrelated to company fundamentals (eg. paying taxes or purchasing a new house). To assess performance, we compare insider-driven stock returns to those of the SPY ETF, a widely used fund that mirrors the S&P 500. We evaluate performance over the six months following each insider purchase.

## Hypothesis

We propose that most insider purchases, while suggestive of insider optimism, will not consistently outperform the market. However, **we hypothesize that a small subset of insiders, with deeper insight and good timing, will generate significant alpha, defined as returns at least 20% higher than the market over the following six months.**

## Caveats

While our finalized dataset is large, spanning thousands of companies and over 100,000 insider transactions, it does not represent the entire universe of insider activity. One important limitation is survivorship bias because our analysis only includes companies that remain publicly traded today. Firms that failed were delisted from the exchange and are excluded (eg. during the 2008 financial crisis). This may skew results toward more successful companies and overstate average performance.

# Data Sources

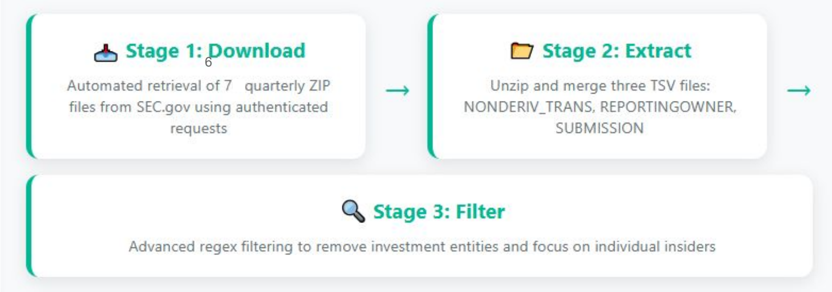
## Primary Dataset

Our primary dataset consists of SEC Form 4 insider trading filings, sourced directly from the Securities and Exchange Commission's publicly available archive (https://www.sec.gov/data-research/sec-markets-data/insider-transactions-data-sets). This comprehensive dataset spans Q1 2006 through Q1 2025, capturing nearly two decades of insider trading activity across all U.S. public companies.

**~305,000**
Insider transactions

**~7,200**
Unique companies

## Data Acquisition Pipeline



**Stage 1: Download**
Automated retrieval of 76 quarterly ZIP files from SEC.gov using authenticated requests

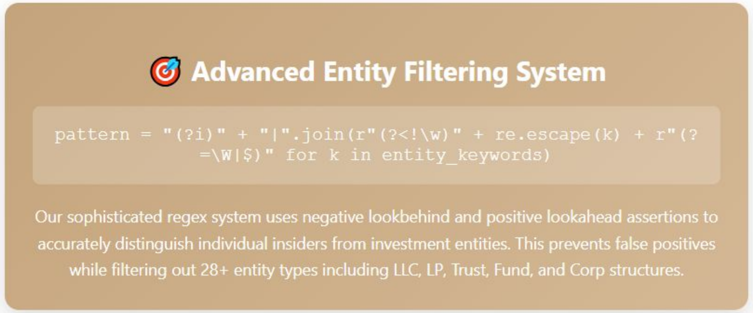
→

**Stage 2: Extract**
Unzip and merge three TSV files: NONDERIV_TRANS, REPORTINGOWNER, SUBMISSION

→

**Stage 3: Filter**
Advanced regex filtering to remove investment entities and focus on individual insiders

We implemented a systematic 3-stage approach to ensure reproducible coverage and data quality. Step 1 automates the download of 76 quarterly zip files. Step 2 extracts and merges three critical TSV files using accession numbers as primary keys. Finally, Stage 3 applies advanced filtering via Regular Expressions (Regex) to distinguish individual insiders from investment entities.

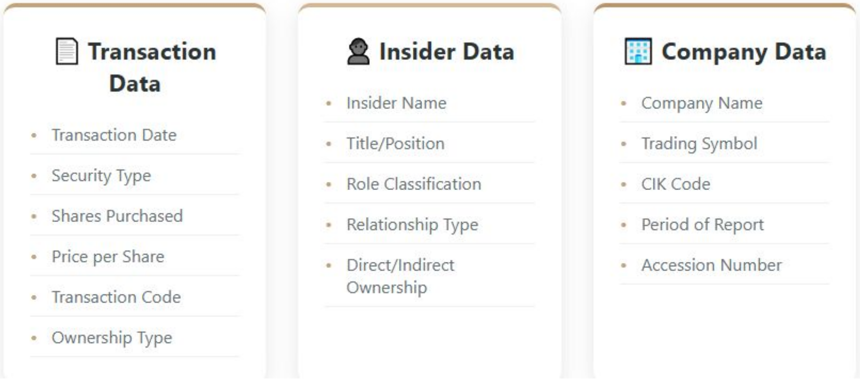


**Advanced Entity Filtering System**

```
pattern = "(?i)" + "|".join(r"(?<!\w)" + re.escape(k) + r"(?
=\W|$)" for k in entity_keywords)
```

Our sophisticated regex system uses negative lookbehind and positive lookahead assertions to accurately distinguish individual insiders from investment entities. This prevents false positives while filtering out 28+ entity types including LLC, LP, Trust, Fund, and Corp structures.

Our most innovative contribution is the regex-based entity filtering system. Using negative lookbehind and positive lookahead assertions, we accurately identify and remove 28+ types of investment entities (LLC, LP, Trust, Fund, etc.) while preserving legitimate individual insider transactions. This prevents false positives like 'PRINCETON JOHN' being flagged as 'INC' while correctly filtering investment entities like 'SMITH INVESTMENT LLC'."



**Transaction Data**
- Transaction Date
- Security Type
- Shares Purchased
- Price per Share
- Transaction Code
- Ownership Type

**Insider Data**
- Insider Name
- Title/Position
- Role Classification
- Relationship Type
- Direct/Indirect Ownership

**Company Data**
- Company Name
- Trading Symbol
- CIK Code
- Period of Report
- Accession Number

Each insider transaction captures detailed information across three key dimensions. The result is a clean, verified dataset of corporate insider transactions, eliminating institutional noise and focusing exclusively on genuine insider trading signals.

# Data Sources

## Secondary Dataset

After filtering our first dataset, we were able to secure data for approximately 305,000 transactions between Q1 2006 and Q1 2025. From this data, we had 7,201 unique stocks that we used to query the yahoo finance API, ‘yfinance’. This API is a popular open source library for python that reliably scrapes the yahoo finance website for corporate information and historical price data which is primarily intended for education and research purposes (https://ranaroussi.github.io/yfinance/).

CSV
Query for Data
yfinance

**Corporate Data**

- Sector
- Industry
- Market Cap
- Beta
- PE Ratios
- Dividend Yield
- 52wk High/Low

**Price Data**

- Date
- Open
- High
- Low
- Close
- Volume

**Calculated Data**

- Percent Change
- Moving Average
- Rolling Trend of Moving Average

**~42%**

**Of companies were listed on Yahoo! Finance**

**~100,500**

**Insider transactions combined with price data**

**~1,600**

**Transactions occurring in the last 6 months**

## Data Aggregation

Our financial data aggregation process consisted of three primary steps. First, for each unique stock symbol, we queried the ‘yfinance’ API to retrieve all daily transaction data, including open, high, low, close prices, and daily trading volume. We then calculated a 28–period moving average to smooth the time series and applied a second 28-period moving average to extract the longer-term monthly trend of price movement.

Next, for each insider transaction associated with a given stock symbol, we collected corresponding price and trend data for the month prior to the transaction, the transaction date itself, and at monthly intervals for six months following the transaction. We also retrieved historical price data for our benchmark, the SPY ETF, and aligned it with the same transaction timeline to enable direct comparisons.

In total, we successfully obtained price data for approximately 42% of our unique stock symbols and over 100,000+ insider transactions. Based on ‘yfinance’ documentation, common reasons for failed requests include invalid or delisted symbols (the most frequent issue), missing exchange suffixes for international securities (not applicable), limited historical data, rate limiting (addressed) and unsupported asset classes (not applicable). Given the nature of our dataset, most failures likely stem from delisted stocks.

# Data Manipulation and Exploratory Analysis

## Form 4 - Transaction Dates and Prices

Form 4 data filed with the SEC is not free of errors, so it was essential to validate the accuracy of the reported transactions before analysis. To ensure data quality, we first grouped transactions by insider and date, then evaluated the consistency of transaction dates. Since our dataset spans from 2006 through 2025, we excluded any entries with dates falling outside this range **(Fig 2a)**.

Additionally, we identified a number of records with stock prices that appeared inconsistent or abnormal **(Fig 2b)** to the price on the date of reported transaction. To address this, we compared the reported average price per share for each transaction against the asset's actual trading range (daily high and low) on the transaction date. Any transactions with prices falling outside this range were flagged as inaccurate and were removed to maintain the integrity of the dataset.

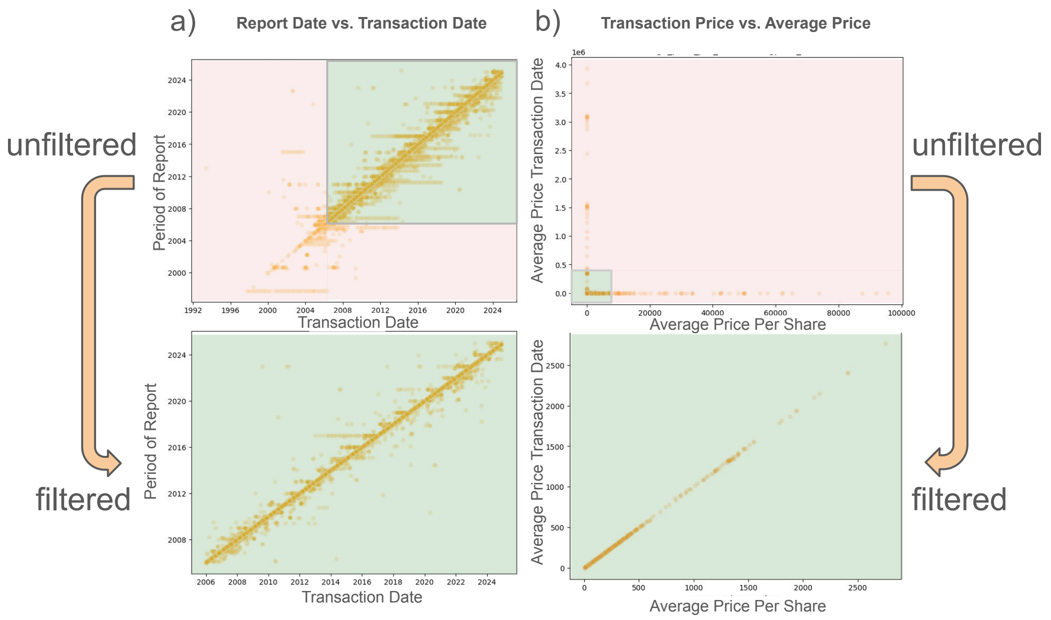


**Figure 2:** Visual representation of data filtering using a) date and b) transaction price. Green blocks are the data that was used for further analysis.

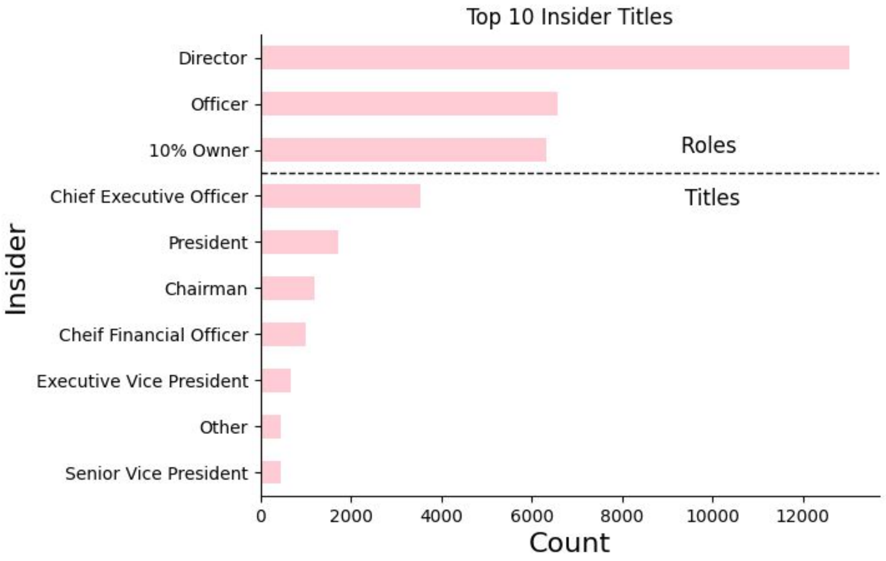


**Figure 3:** Top 10 insider roles and titles. Roles refer to generalized categories of insider positions, while titles represent the specific job titles held by individuals.

## Form 4 - Insider Roles and Titles

Insider roles and titles reported on Form 4 lack a standardized format, and individuals often hold multiple positions simultaneously. As a result, this field exhibited considerable variability and required manual normalization for meaningful analysis. We systematically cleaned and standardized the reported roles and titles to ensure consistency across records, enabling us to analyze insider activity based on position more effectively.

Insiders primarily fell into one of three role categories: Director, Officer, or 10% Owner. Among these, Directors were the most active, accounting for over 13,000 transactions in our dataset. Titles among Officers varied widely, though the most frequently observed was Chief Executive Officer, with just over 3,500 transactions recorded **(Fig 3)**.

# Data Manipulation and Exploratory Analysis

## Yahoo! Finance Price Data

As part of our data cleaning and exploratory data analysis, we employed a scatter plot matrix to visually inspect the aggregated price data and identify any potential anomalies. Yahoo! Finance is known for it's clean and accurate data but we wanted to confirm this ourselves.

Our initial analysis focused on all individual stock data **(Fig 4a)**, followed by a separate examination of the SPY ETF price data **(Fig 4b)**. The distribution of 'stock' prices exhibited a pronounced right skew, with the majority of prices falling below $500/share. While there was natural variation in prices over the seven month period, we did not observe any extreme outliers.

In the case of the SPY ETF, price data spanning the past two decades revealed a wide range of values, consistent with long-term market growth. Notably, we observed heightened volatility around the $450/share mark, potentially reflecting periods of increased market uncertainty.

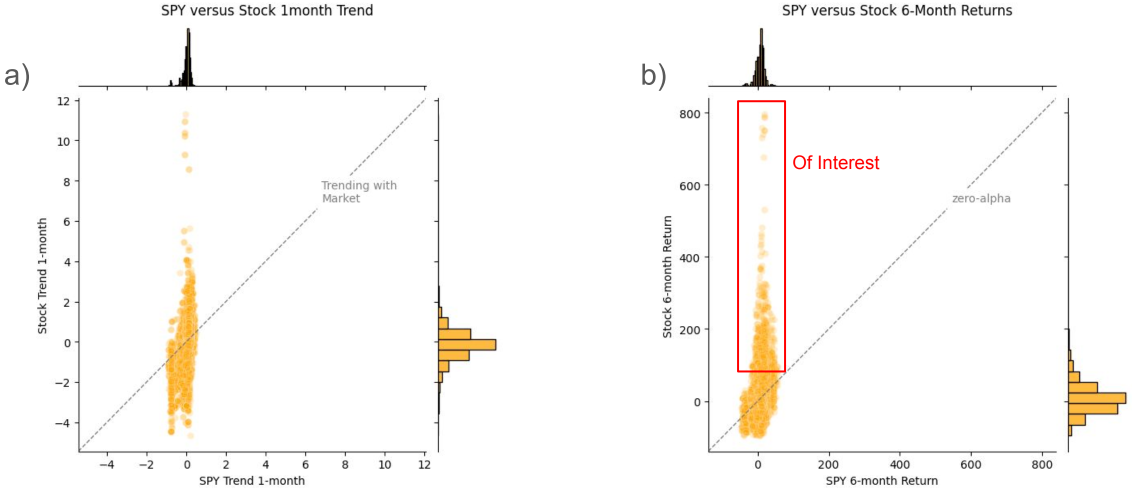


**Figure 5:** Joint plot a) Showing stock and SPY ETF trend for the month following insider trading activity. b) Comparing absolute returns of holding the stock versus the SPY ETF after six months. Anything above the diagonal lines shows market outperformance.

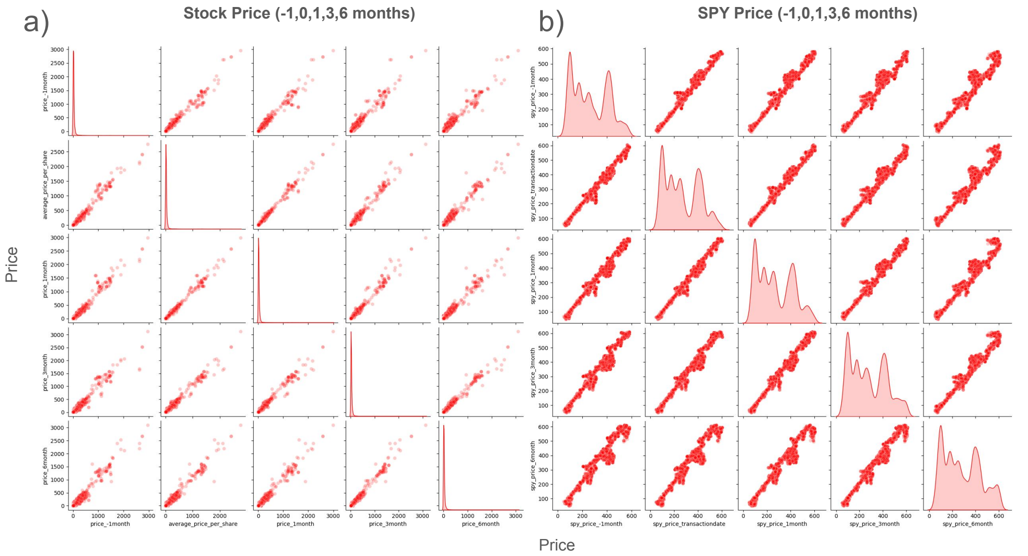


**Figure 4:** Scatter Plot Matrix comparing a) stock and b) SPY prices. Comparisons are made over a seven month period from 1 month before the transaction to six months after. This is an easy visual representation to identify any potential outliers.

## Calculated Trends and Returns

Our original hypothesis posited that the majority of insider transactions do not outperform the market and are effectively noise. To explore this, our initial exploration needed to compare all insider transactions to the market benchmark (SPY), focusing on both price trends and return behavior **(Fig 5)**. We calculated returns using the standard formula $r_t = P_{t2} - P_{t1} / P_{t1}$, where $P_{t1}$ is the previous price and $P_{t2}$ is the price at the comparison point.

The distribution of returns exhibited a significant right skew **(Fig 5b)**, a common characteristic in financial return data. Most returns clustered near zero. Further inspection revealed no extreme outliers attributable to data errors, suggesting the return patterns observed were genuine and not artifacts of faulty inputs.

# Insider Position Analysis

## Distribution of Returns: Insiders vs. Market

Due to the right-skewed nature of the return distribution, we employed the Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test, to evaluate our hypothesis.

- **Null Hypothesis ($H_0$):** Insider transactions yield returns similar to the market, indicating no informative value.
- **Alternative Hypothesis ($H_a$):** Insider transactions either outperform or underperform the broader market.

The test results indicated that the **median return** from insider transactions was significantly lower than that of the market ($p<0.05$), providing statistical evidence that insiders, on average, underperformed the benchmark **(Fig 6)**.

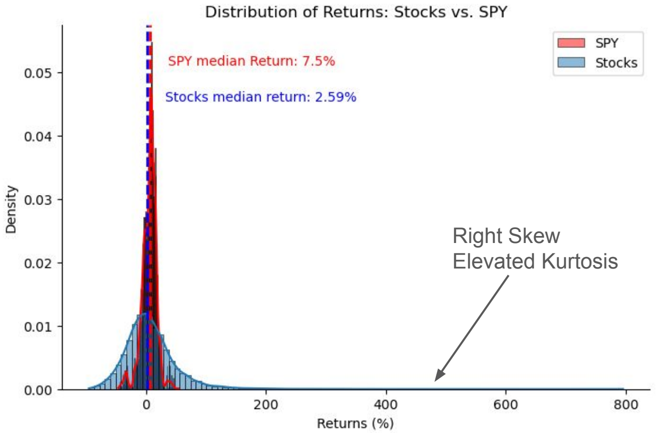


**Figure 6:** Distribution of returns for 6 months following insider transactions. Returns are compared to the SPY ETF. The median return is reported as the primary measure of central tendency for non-parametric data.

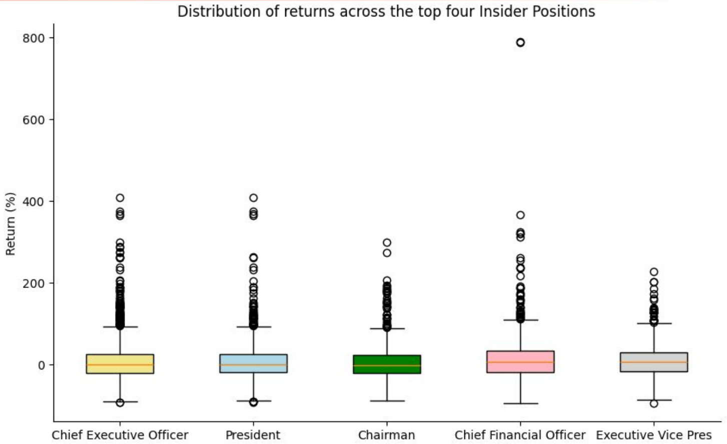


**Figure 7:** Box plot of returns for 6 months following insider transactions based on insider position. The median return is reported as the primary measure of central tendency for non-parametric data.

## Performance Comparison by Insider Position

To explore whether certain insider roles are associated with superior returns, we conducted a Mann-Whitney U test, a non-parametric method appropriate for comparing independent groups.

- **Null Hypothesis ($H_0$):** There is no difference in returns based on insider role.
- **Alternative Hypothesis ($H_a$):** Insider roles with privileged access to corporate financials, such as the Chief Financial Officer, may generate higher returns than others.

The test revealed **no statistically significant differences** among the median returns of the top five insider positions **(Fig 7)**, suggesting that no single role consistently outperformed others in terms of post-transaction returns.

## Do Insider Positions Predict Alpha?

Taken together, these results suggest that insider transactions alone are not reliable predictors of alpha. Most trades cluster around zero excess return and monthly price trend, reinforcing the hypothesis that the majority of insider activity is random noise **(Fig 8)**. However, the right-skewed distribution and elevated kurtosis indicate the presence of occasional, outsized returns. This suggests potential for identifying alpha-generating trades using additional contextual behavioral variables beyond role.

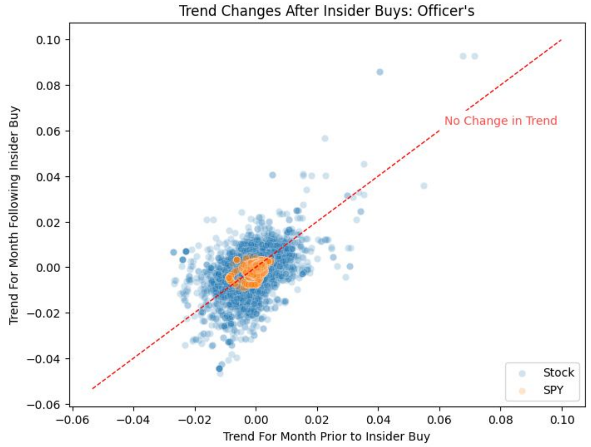


**Figure 8:** Distribution of average stock trend per day in the month following Officer insider transactions. While insider-related stocks exhibited greater volatility, their overall trend closely mirrored that of the broader market.

# Individual Insider Analysis

## Elite Alpha Insiders: The Key to Unlocking Value from Insider Activity

While our positional analysis found no consistent relationship between job title and superior trading returns, a deeper dive at the individual level revealed a remarkable insight: A small subset of elite insiders that generated substantial alpha. After applying quality filters, we identified **388 "elite alpha insiders"** –each with 3 or more buys that yielded raw 6-month excess returns ≥ 20% (i.e., stock return less SPY return).

**Figure 9** displays the distribution of both raw and beta-adjusted excess returns for these insiders:

- Raw excess returns were clustered in the 30%-50% range, indicating consistently strong performance.
- Beta-adjusted excess returns, which normalize for risk by dividing raw excess return by the stock's volatility (beta), showed a flatter and more dispersed distribution–as expected–but still retained a median of ~40%, underscoring robust alpha generation even after accounting for volatility.

These results suggest that some insiders are acting on a combination of informational advantage and market-timing skill.

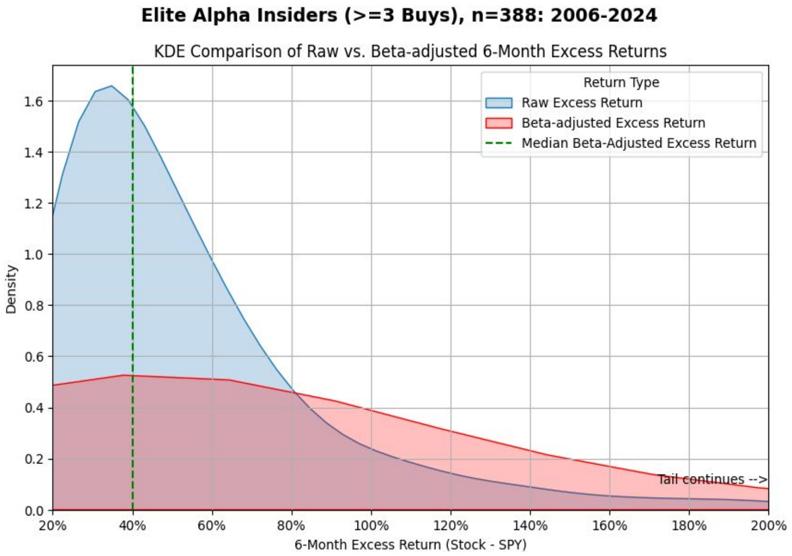


**Figure 9:** There are insider buyers who identify mispriced stocks.

## Sector Concentration of Elite Alpha Insiders

**Figure 10** reveals the sector distribution of insider buyers who met the "elite alpha" criteria.

**Healthcare** dominates with more than 36% of insider alpha trades, more than double the next sector. This suggests strong informational advantages or market inefficiencies in this space, possibly due to complex product pipelines and asymmetric access to clinical or regulatory information.

**Technology, Industrials, and Consumer Cyclicals** collectively account for almost 47% of elite alpha buys, reflecting dynamic and event-driven sectors where insider-timing may be more valuable.

**Utilities, Basic Materials and Real Estate** had few elite alpha buys, reflecting slower-moving and more transparent industries where alpha opportunities from insider knowledge may be more limited.

This concentration reinforces the idea that insiders' information advantage is strongest in sectors where uncertainty and asymmetric information are highest.

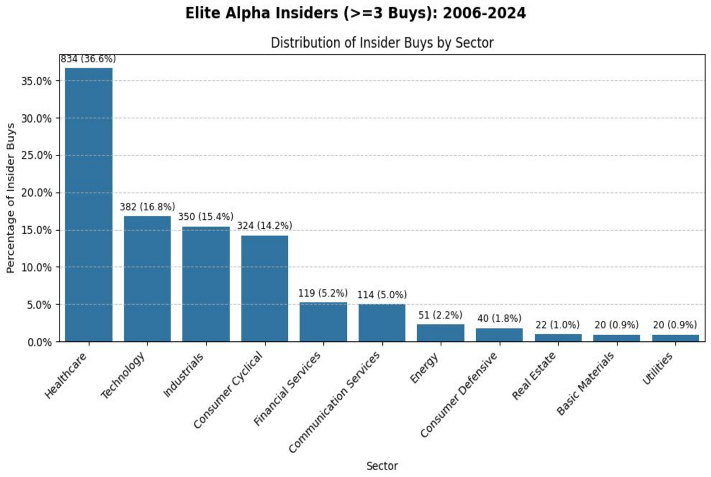


**Figure 10:** Sector concentrations reflect information advantage

# Conclusions and Future Work

## Sift Through The Noise To Find The Gold

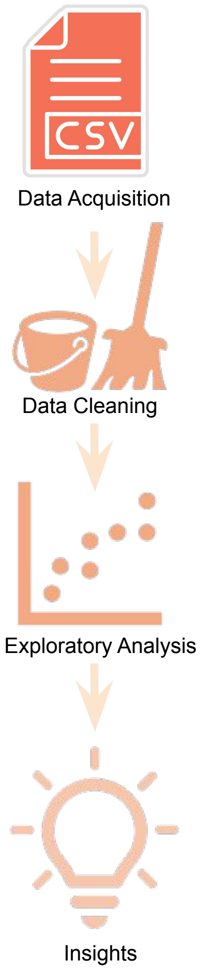
The U.S. Securities and Exchange Commision (SEC) requires corporate insiders to disclose stock acquisitions or dispositions in their own firms within 48 hours via Form 4 filings. These mandated disclosures offer a valuable opportunity to extract unstructured data that can inform financial models designed to outperform the broader market and generate alpha.

In this study, we constructed a novel dataset by merging SEC Form 4 data with market and corporate information from the Yahoo! Finance API (yfinance). The result was a comprehensive dataset covering over 3,000 companies and more than 100,000 insider transactions, enabling detailed analysis at both the role and individual level.

Throughout the data engineering process, we addressed challenges such as inconsistencies in transaction dates, stock prices, and insider role classifications. Using data science techniques such as regular expressions, normalization, and robust filtering, we were able to deploy a clean and reliable dataset to support our analysis.

Our results indicate that, in aggregate, insider trading activity behaves like random noise and often underperforms the market. However, we identified a distinct subset of individuals that we term **'elite alpha insiders' who consistently generated six-month, risk adjusted returns exceeding 20% above and beyond the market benchmark.** These finding suggest that while most insider activity lacks predictive value, there exists a small but potentially exploitable signal within the noise.

## Data Science Pipeline



Data Acquisition

Data Cleaning

Exploratory Analysis

Insights

## Future Work

As an extension of this project, future research should explore alternative resources for stock data to overcome the limitations encountered with the Yahoo! Finance API. Both free and paid platforms may offer enhanced coverage, reliability, or granularity:

- **Alpha Vantage:** Comprehensive financial API (Free and Paid)
- **Finnhub:** Institutional-grade financial data (Free and Paid)
- **Marketstack:** Specializes in historical market data (Paid)
- **Alpaca:** Commission-free brokerage with robust API (Account Required)
- **Tiingo:** Real-time and historical price data (Paid)

Beyond expanding data sources, further research could deepen the analysis of successful insider trading patterns by incorporating dimensions such as sector classification, market capitalization, and macro-level market trends.

The enriched dataset produced in this project offers a strong foundation for integration into a machine learning pipeline. Both unsupervised (e.g., clustering insider behavior) and supervised learning (e.g., forecasting post-trade performance) can be explored to uncover predictive relationships and enhance model-driven investment strategies.

# Statement of Work and References

## Statement of Work

Collaboration raised unique challenges. We gained experience in collaborative tools like Google Colab and GitHub. Our team had multiple video meetings and communicated through slack extensively. In the future, we can improve collaboration by more clearly outlining project tasks for each individual and creating an environment where colleagues are encouraged to engage and ask questions.

**Thomas MacPherson**

Project proposal, financial data aggregation, data manipulation, exploratory data analysis, Insider role and position analysis, visualizations, report writing and editing

**Kirtland Corregan**

Project proposal, Insider data acquisition, data manipulation, exploratory analysis, Individual Insider analysis, visualizations, report writing and editing

**Rami Abushamalah**

Editing project proposal, Corporate data acquisition, exploratory analysis, visualizations, report writing and editing

## References

- Hussain, A. (2022, December 09) *Insider: Definition, Types, Trading Laws, Examples*. Investopedia. Retrieved from https://www.investopedia.com/terms/i/insider.asp

- Ran, A. (n.d.) y*finance: Yahoo! Finance market data downloader.* GitHub. Retrieved from https://ranaroussi.github.io/yfinance/

- Charles Schwab & Co., Inc. (2023, November 20). *How well do you know market cap?* Charles Schwab. Retrieved from https://www.schwab.com/learn/story/how-well-do-you-know-market-cap

- U.S. Securities and Exchange Commission. (n.d.). *Insider Transactions Data Sets*. SEC.gov. Retrieved from https://www.sec.gov/data-research/sec-markets-data/insider-transactions-data-sets

- U.S. Securities and Exchange Commission. (n.d.). *Insider Transactions ReadMe* [PDF file]. Retrieved from https://www.sec.gov/data-research/sec-markets-data/insider-transactions-data-sets