SIADS 593: Milestone I

Team Project Proposal

version 2025.05.13

**Instructions:** please make a ***copy*** of this template file (do not edit original).

## **Proposal Title:** Insider Trading: Do Corporate Insiders Know Something We Don't?

## 1. Team members

Please list your team members (2-3 max).

* Kirtland Corregan: [kirtland@umich.edu](mailto:kirtland@umich.edu)
* Rami Abushamalah: [ramiab@umich.edu](mailto:ramiab@umich.edu)
* Thomas MacPherson: [Thomas Macpherson](mailto:tmacphe@umich.edu)

## 2. Project summary

Summarize your proposed project in a few sentences.

#### What is your proposed project and why are you proposing it?

#### What are the question(s) you want to answer, or goal you want to achieve?

| We propose a team project that is looking into corporate insider transactions as a means to predicting the general trend of a stock over the following 3 to 6 months. According to our research, insider trading has long been used to gain an edge in the market. However, we want to take the analysis one step further with the skills we have gained so far in the Masters of Applied Data Science program at the University of Michigan. More precisely, we want to see if we can identify insiders or executive positions that tend to have better predictive accuracy compared to others. Some executive positions, such as the Chief Executive Officer, tend to have a more general positive disposition towards companies and may come from more of a marketing or business background. We hypothesize that executive positions such as the Chief Financial Officer or Chief Accounting Officer have a better grasp on the general health of the company and if they are buying, we should be buying as well. |
| --- |

## 3. Datasets

#### Describe one primary dataset and at least one secondary dataset. If other secondary datasets will be used please describe them as well.

#### The proposed datasets should exhibit different features/columns and/or different access methods, e.g., \*.csv file, \*.json file, API retrieval, web scraping, etc. Different time periods, for example, with the same features/columns is not considered a different dataset. Remember, the focus of the project in this Milestone course is to give you the opportunity to practice your data manipulation skills, so feel free to challenge yourself.

#### If you're unsure if your data sets are "different enough" describe the datasets and request a review via the *#siads593\_[semester]\_001\_project* Slack channel.

#### **Please note:** all proposed datasets ***MUST*** be publicly available to all members of the class (students, instructors, course support personnel, etc.). Use of proprietary datasets for this project is ***not*** permitted.

## 3.1 Primary dataset description

Describe your primary dataset. How is the data collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

| Our primary dataset will be archived, historical insider data files downloaded from the SEC (<https://www.sec.gov/data-research/sec-markets-data/insider-transactions-data-sets>). These .zip-format files are grouped by calendar quarter, with each ZIP usually containing ten files.  We will focus on three categories of the files, which are in TSV format (tab-separated values): ‘NONDERIV\_TRANS.tsv’, ‘REPORTINGOWNER.tsv’, and ‘SUBMISSION.tsv’. This file format can be processed similarly to a CSV file.  These files contain important data fields, such as: Name of the insider, the insider’s title (role), the issuer company, the company’s ticker symbol, the period of report, the transaction date, the security title (e.g. “common stock”), the transaction code (for buy or sell), the number of shares, the price per share, the total shares owned, the ownership type, and the accession numbers. We will be obtaining all records from 2006 to Q1 2025.  With respect to data size/volume, the quarterly archived ZIP files range in size from 7 MB TO 15 MB. Currently, there are 19 years and 1 calendar quarter of data, representing 77 total quarters of archived data at the SEC link highlighted above. We estimate approximately 850 MB of raw data being processed. We estimate that filtering and cleaning will reduce this raw data to approximately 20 MB, prior to further analysis.  Total insider trade records per month are typically in the single-digit thousands range, with variability. The total volume of insider activity varies from quarter to quarter, depending in part on financial market conditions, as well as the seasonality of holidays and quarterly earnings reports, with not all companies being on a standard calendar quarter. Further, given that our project is focused on “open market” insider buying, rather than both buying and selling, there will be more variability in the volume of quarter-to-quarter insider trading records. For example, if there has been a major market sell-off due to macro concerns, there may be an increase in insider buying if insiders perceive better value in stock prices. By “open market,” we mean common stock purchases made by insiders at prevailing market prices, not option-related transactions, which can occur outside of market prices.  In terms of access method, SEC Form 4 data is public, structured and accessible via direct HTTP links. Accordingly, our code will download the archived zip files via HTTP GET requests:  (Example: <https://www.sec.gov/files/structureddata/data/insider-transactions-data-sets/2025q1_form345.zip> ).  Our code then loads multiple quarterly ZIP files, and extracts and processes the TSV files inside each archive. The TSV files are merged/joined using “ACCESSION\_NUMBER” as the join key. The code then filters insider transactions to focus on open-market purchases by individual insiders who are officers and directors of the company. It further filters out purchases by investment entities such as funds, LPs and trusts. The processed results are compiled into a DataFrame and saved to a CSV for backup and potential upload to a database or machine-learning pipeline, where this data and analytical results could be part of a multi-factor investment strategy. |
| --- |

## 3.2 Secondary dataset(s) description

Describe your secondary dataset(s). How is the data collected and how will you access it? Please share what features in the dataset(s) are relevant to your topic and describe the data types you’re expecting. At a minimum, for each secondary dataset include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

| Our second dataset will be obtained using the yahoo finance API, ‘yfinance’. The yahoo finance API is a popular open source library that reliably scrapes the yahoo finance website for corporate information and historical price data which is primarily intended for education and research. You can find documentation on yfinance here: ([**https://ranaroussi.github.io/yfinance/**](https://ranaroussi.github.io/yfinance/)**).**  We will be using information from our first dataset (transaction date, ticker symbol) to query the API for corporate information as well as historical price data.  The data related to the company will include:   * Current Market Cap * Sector * Industry   We will use the insider transaction dates to obtain historical price data surrounding the time of the insider transaction (1 month before transaction to 6 months after transaction). Each of these data frames will contain approximately 128 rows and 6 columns of daily price data (open, high, low, close, volume). We will be performing this for every unique transaction in the insider trading dataset which will aggregate to a massive amount of data. The yahoo finance API allows us to directly import using pandas in order to manipulate and organize this data in a format that we can ultimately merge with our originally constructed dataset. This merge will be based on the date of insider transaction and ticker symbol thus creating a completely novel dataset that we can use for data exploration. |
| --- |

## 3.3 [ Yes ] Affirm: datasets are public.

Please write YES in the above box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

## 4. Cleaning and manipulation

Describe how you will need to manipulate your datasets: how will you handle missing or anomalous data? How will you join your primary and secondary datasets? What cleaning and manipulation challenges, if any, do you anticipate?

| Original Dataset:   * Obtaining pertinent information * Filtering based on secuirtyTitles (common Stock) * Filtering by transaction codes (only looking at buy side transactions) * Merging files based on accession numbers * Exporting to CSV/database files * Groupby transactions made in short timeframe by same individuals * Clean titles, roles, remove missing tickers, look for abnormal data/outliers   Secondary Dataset:   * Query yahoo finance with dynamic variables from original dataset * Combine a standard historical price object for comparison (SPY ETF) * Work with datetime objects in pandas * Remove missing values * Confirm proper industry/sector, SIC codes * Use timestamps and timedeltas to pull out specific historical price data * Vectorized calculations of moving averages for general stock trends   Novel Dataset:   * Vectorized mathematical operations for novel features (standardized price change based on SPY) * Grouping by insider, insider role, industry, size of transaction, look for novel patterns in the data |
| --- |

## 5. Analysis

Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

| Our primary analysis will be looking at whether the correlation of CFO purchases vs stock gain is more significant than other corporate positions. We will further analyze whether particular individuals have a better track record than others in that role. We will compare the correlation to a market mimicking ETF like the SPY (S&P 500) to see whether the stock is moving due to general market trends or the strength of the individual company. |
| --- |

## 6. Visualizations

Describe in 1-3 sentences at least **two** data visualizations that you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

| Visualization 1 - Line chart showing general price trend for a given ticker before and after insider transaction based on end of day close prices  Visualization 2 - Bar Chart to analyze how active different positions are in buying company stock. Who is the most active  Visualization 3 - Scatterplots for correlation based on size of purchase and standardized price changes in stock  Visualization 4 - Screen shot of the interactive dashboard being hosted on a webpage |
| --- |

## 

## 7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

| Yes, we think all studies have some ethical implications. In our case, it will be imperative to keep the endpoint stakeholders in mind. People may use this information to invest hard earned money, whether it is their retirement fund, kids education fund, or money set aside for a rainy day. It is our ethical/moral obligation to make sure that we adequately communicate the risk and uncertainty that goes along with our analysis.  Our analysis will not directly buy or sell shares of any companies, so ultimately a human will be involved in any financial risk decisions. |
| --- |

## 8. Contributions

Indicate the contribution that each team member will make to the project.

| Our goal is to truly work on this project as a team through each step. This means each member being involved in every step of the process. Not only will this help each of us in our understanding of the concepts, coding, and analysis but it will bring different perspectives to each step. We want to avoid the ‘divide and conquer’ approach as it may leave the final project feeling clunky. Kirt brings in years of valuable experience from the financial sector, Rami brings in extensive computer coding and dashboard building skills. Thomas bridges the gap between mathematics, finance, and coding. Together, we think we make a very strong team.  Project Proposal   * Thomas, Kirt, Rami   Data Acquisition   * Kirt, Thomas, Rami   Data manipulation and cleaning   * Rami, Kirt, Thomas   EDA and Visualizations   * Rami, Thomas, Kirt   Analysis and Conclusions   * Kirt, Thomas, Rami   Report Construction   * Thomas, Kirt, Rami   Interactive Dashboard   * Rami, Kirt, Thomas |
| --- |

### 

## Changelog

(2025.05.11 TM) Initialization of the document, filled out title, filled out group members, rough draft on sections.

(2025.05.12 KC) Edited and updated Section 3.1 “Primary dataset description.”

(2025.05.13 KC) Edited and updated Section 3.1 “Primary dataset description.”

(2025.05.13 TM) Edited and updated proposal for clarity