Data Management Project

Leonardo Palestra, Marco Gelsomini
Università degli studi Milano-Bicocca, CdLM Data Science

Abstract
This project aim to explore the relationships between the stock market and the transactions of the insiders (which are every director, officer and owner of more than 10 percent of a class of a particular company's equity). After illustrating the processes of collecting, storing and integrating the data, an EDA is reported. The base question is whether the insiders can make their decisions of buying or selling anticipating a favorable move in the stock price. The SEC offers a pre-defined automatic buying and selling plans (10b5-1 Plan) that insider can adopt to avoid future obligations of trading with material non-public information (MNPI). Also, companies typically set a period where the insiders cannot make transactions (for example, near earnings periods). Nevertheless, in the end, they can mantain a certain degree of flexibility.

The result show that there is some evidence of relationship between the stock market and the transactions of the insiders. However, there is the need to inspect the single company in order to understand the single case and to be able to explain how this relationship works.

Contents

1 Data Sources

To conduct a thorough analysis, we collected data from three primary sources: the NASDAQ official website, the Alpha Vantage API, and the SEC website via web scraping. The entire data retrieval process was automated using Python.

1.1 NASDAQ Official Website

The first step in our data collection process was obtaining the list of all companies listed on the NASDAQ stock exchange. We accessed the official NASDAQ website and extracted company names, ticker symbols, and market capitalization values. Since our study focuses on the most significant companies, we filtered the list to retain only the top 50 firms by market capitalization. This filtered list was then used as the basis for further data collection.

Data Extraction and Filtering Process:

1. Access the NASDAQ official website.
2. Retrieve the full list of listed companies, including ticker symbols and market capitalization.
3. Use Python to filter and retain only the top 50 companies by market capitalization.
4. Store the filtered list in memory for subsequent API calls and web scraping.

1.2 Alpha Vantage API

With the refined list of 50 NASDAQ-listed companies, we used the Alpha Vantage API to collect daily stock price data for each of them. The API provided key financial metrics, including opening price, closing price, daily high and low prices, and trading volume. To have a broader perimeter in this initial step we collect the last 20 years of data.

Data Retrieval Process:

1. Register for an API key on the Alpha Vantage platform.
2. Iterate through the list of 50 selected companies, sending API requests for daily stock prices.
3. Retrieve and parse the JSON responses, extracting and elaborating relevant financial data.
4. Store the stock price data in memory for further analysis.

1.3 SEC Website (Web Scraping)

To analyze insider trading activity, we extracted insider transaction reports from the official website of the U.S. Securities and Exchange Commission (SEC). We automated the data collection using web scraping techniques.

Web Scraping Process:

1. Iterate automatically through the list of 50 selected companies, performing web scraping for each one, dealing also with the automatic change of all the available pages.
2. Extract in json format all raw information from the XML document associated with each form-4 filling. This form is requested within 2 days when an insider executes a transaction.
3. Store the extracted transaction data in memory to obtain historical data for further processing.

Potentially the scraper can be launched every time to obtain new data by choosing a proper date window.

2 Storage

The collected data was stored in MongoDB, separated into two collections:

- **Stock**: Contains daily stock price data for each company, retrieved via the Alpha Vantage API.
- **Insider_trading**: Stores insider trading transactions extracted from SEC filings via web scraping.

2.1 NASDAQ Company List

The NASDAQ company list was used to define the scope of the analysis. Initially, all companies listed on the NASDAQ were retrieved from the official website. Using a Python script, we filtered the dataset to retain only the top 50 companies by market capitalization. This final list was then used to fetch stock prices and insider trading data. The list itself was not stored in the database but was kept in memory during data retrieval.

2.2 Stock Price Data Storage

For each of the selected 50 companies, we queried the Alpha Vantage API to fetch historical stock prices.

The raw API response contained unnecessary fields and inconsistent field names, so before storing the data, the following transformations were applied:

- Renamed API field names to be more MongoDB-friendly (e.g., `"1. close"` → `"close"`).
- Create a new field with the converted dates to ISO date format.
- Organized data so that each stock has a single document containing an array of daily price records. The array make it easier to retrive informations and reduce the memory required. The timestamps that were previously keys are now a field inside each daily data.

Document structure (Stock Collection)

```
{
  "Meta Data": { "Symbol": "AAPL", "Last_Refreshed": "2025-02-14" },
  "Time Series (Daily)": [
    { "date": "2025-02-14", "open": 241.25, "close": 244.60, "isodate": ISODate(...) },
    { "date": "2025-02-13", "open": 236.91, "close": 241.53, "isodate": ISODate(...) }
  ]
}
```

Each stock has exactly one document in the collection, with an array storing multiple daily price records. Given that we stored only 50 stock documents (each containing multiple time series entries), the data volume remains manageable.

2.3 Insider Trading Data Storage

For insider trading transactions, we performed web scraping on the SEC website, cycling through the list of selected companies. Each SEC filing retrieved was processed to extract relevant fields, including:

- Insider's name and company and its role inside the organization.
- Transaction details (type, date, number of shares, price per share, buy or sell).
- Filling metadata (issuer, reporting owner, ownership type).

Each document in the `Insider_trading` collection corresponds to an individual SEC filing compiled by a single entity. Inside a single document more transaction can be registered for one person on the same transaction date. Moreover multiple documents by different entities can referred to the same transaction date. This resulted in a total of 18,914 documents in this collection. The structure of the document is complex in this raw format due to the fact that data are stored directly without processing from the XML the produce this type of document.

Document structure (Insider Trading Collection)

```
{  "id_":   ...,
  "Date_transactions": "2020-04-28"
  "Date_publication": "2020-04-30",
  "Company": "AAPL",
  "Filling_data": {
    "issuer": { "issuerName": "Apple Inc." },
    "reportingOwner": { "rptOwnerName": "JUNG ANDREA" },
    ...
    "nonDerivativeTable:{"nonDerivativeTransaction":[...]}
    ...
    ...
  }
}
```

2.4 Why MongoDB?

MongoDB was chosen as the storage solution for several reasons:

- **JSON-based Data Structure**: Both the Alpha Vantage API and the SEC filings provide data in JSON format, which aligns naturally with MongoDB's document model. This eliminates the need for complex data transformations before storage.
- **Scalability for Large Datasets**: The insider trading dataset consists of 18,914 documents, whereas the stock price dataset consists of only 50 documents but with longer arrays inside. MongoDB efficiently handles large numbers of documents distributing data across multiple nodes while allowing flexible indexing and querying.
- **Flexible Schema**: Insider trading data varies in structure depending on the filing type. Unlike relational databases, MongoDB allows us to store such heterogeneous data without enforcing a rigid schema.
- **Efficient Time-Series Queries**: Stock price data follows a time-series pattern, which can be efficiently queried in MongoDB using date-based indexes.
- **Best for Integration**: The final analysis and integration of the data will also be performed within a MongoDB environment, making it the most efficient choice for data storage and retrieval.

This storage setup ensures that both stock price trends and insider trading activity can be analyzed efficiently.

3 Data Integration

3.1 Data pre integration

To integrate the datasets, we performed the following transformations, creating two refined collections: `Stock Performance` and `Insider Trader Cleaned`.

- **Processing the Stock Collection → Stock Performance:** The `Stock` dataset originally contained 20 years of historical date due to API limitations, as Alpha Vantage does not allow selective retrieval of specific date ranges. Since the `Insider Trader` dataset spans only the last five years, we filtered the stock data accordingly to maintain consistency yto match the 5 years span("stock reduce" collection). Moreover we unwind the time series arrays in each document to make each document referring to a single day.

  Additionally, we introduced new fields to measure stock performance over different time horizons. These fields capture percentage changes in stock price and trading volume over 1, 5, and 15 days following the current date. In order to obtained this data we did a "self lookup" operation whit the date of the second collection shifted forward. For the most recent date and for some exceptions such as Friday we set these values to null. In this phase we also convert in number from string values. For example:

```
{"_id": {"$oid": "67b34c5ee1025b242fd2496a" },
  "symbol": "AAPL",
  "date": {
    "$date": "2025-01-16T00:00:00.000Z"},
  "close": 228.26,
  "open": 237.35,
  "high": 238.01,
  "low": 228.03,
  "volume": 71759052,
  "change_close_next_1_day": 0.75,
  "change_close_next_5_days": -2.02,
  "change_close_next_15_days": 2.17,
  "change_volume_next_1_day": -4.56,
  "change_volume_next_5_days": -16.06,
  "change_volume_next_15_days": -58.3}
```

- **Processing the Insider Trader Collection → Insider Trader Cleaned:** The `Insider Trader` dataset was refined into `Insider Trader Cleaned`, retaining only essential attributes relevant to our analysis. First of all we remove the documents that are not belonging to form-4 filling. Then we remove all the documents that have zero data for the non-derivate table as we are interested only in the transactions on the open market. Then we inspect the single transaction within a document and remove the one that has transaction code not related to a simple sell or buy operation (for example tax related, gift,option related transaction...). Finally we clean some string field, convert all the other necessary field and extract the role of the insider. We removed around 10k documents.
- **Creation of indexes:** In order to make query, lookup and other operation we create for both the two sources a composite index that include the date and the company field.
- **Preserving Raw Data for Future Use:** To maintain data integrity, we stored the unprocessed raw datasets in MongoDB. This approach ensures that we can revisit, reprocess, or expand our analysis.

3.2 Integration

To merge the datasets, we performed a lookup operation between `Insider Trader Cleaned` and `Stock Performance`. The integration was based on the two commons key fields: date and company.

This new dataset `Stock Insider Trading integrated` has each document that represents a specific company on a given trading day. Within each document, we included a primary array ”transactions from lookup” containing all insider transactions executed on that date for the respective company. The lookup operation in fact automatically append another document in the top level array as soon as it finds another one matching the key fields. Every elements of this array in the field ”filtered nonDerivativeTransaction” contain another array with the list of transaction or transaction of the same person.

This structure enables efficient analysis of insider trading activities while maintaining the historical stock performance of each company.The composite indexes created after the writing of the two sources help us resolve some memory overloading issues for this set of operations.

3.3 Enrichment

In the final phase, soon after integrating the stock data and insider trading transactions, several additional fields were added to the documents in the `Stock_Insider_Trading_Integrated` collection. These fields provide further insights into the transactions and stock performance:

- `total_entity_transactions`: The total number of transactions per entity.
- `count_acquired`: The number of transactions with the ”A” code (acquisition) per entity.
- `count_disposed`: The number of transactions with the ”D” code (disposition) per entity.
- `total_transactions`: The total number of transactions within the document, summing up the transactions for all entities involved in that date.

These new fields are intended to give a more detailed view of the insider transactions, specifically focusing on the number of acquisitions and dispositions for each insider. The `total_transactions` field provides a summary of all the transactions in a given document, while the `total_entity_transactions`, `count_acquired`, and `count_disposed` fields break down the transactions by individual insiders.

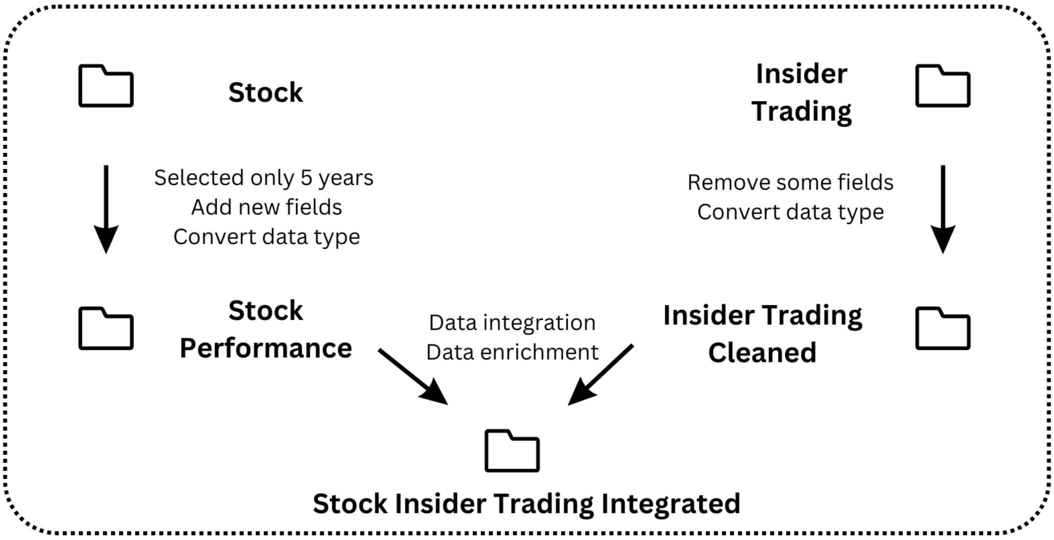
3.4 Final schema

An example of the resulting pseudo-structure is shown below:

```
{
 "_id": {"$oid": "67b34c5ee1025b242fd2522a"},
 "symbol": "NVDA",
 "$date": "2021-04-05T00:00:00.000Z","close": 559.5,"open": 554.7, ... ,
 "change_close_next_1_day": -0.9, ... ,
 "change_volume_next_1_day": -25, ... ,
 "transactions_from_lookup": [
     "$date": "2021-04-05T00:00:00.000Z",
     "filtered_nonDerivativeTransaction": [ 0 [...],1 [...] ],
     "Entity_name_cleaned": "PERRY MARK L",
     "isDirector": "1",
     "isOfficer": "0",
     "isTenPercentOwner": "0",
     "isOther": "0",
     "officerTitle": null,
     "otherText": null,
     "total_entity_transactions": 2,
     "count_acquired": 0,
     "count_disposed": 2,

      "$date": "2021-04-05T00:00:00.000Z",
     "filtered_nonDerivativeTransaction": [ 0 [...],1 [...], 3[...] ],
     "Entity_name_cleaned": "Luca Maestri",
     "isDirector": "0",
     "isOfficer": "1",
     "isTenPercentOwner": "0",
     "isOther": "0",
     "officerTitle": null,
     "otherText": null,
     "total_entity_transactions": 3,
     "count_acquired": 1,
      "count_disposed": 2
  ],
  "total_transactions": 5
}
```

This is the resulting high-level schema of the database (each folder is a collection):



4 Data Quality

In order to evaluate the quality of the collected data, the following metrics were used:

4.1 Data Quality on Stock

- **Temporal Completeness**: This metric focused on checking for any missing dates in the stock price dataset.

  For the 20-year period under analysis, the total number of expected trading dates was 262,484. However, 57,030 dates were found to be missing, which corresponds to approximately 21.7% of the expected dates. These missing dates align with weekends (Saturdays and Sundays) and other holiday, during which the stock market is closed. Therefore, the missing dates are consistent with the non-trading days of the stock market, and no significant data quality issues were detected in this regard.

- **Data Consistency**: This metric checked whether the stock price values were consistent, specifically verifying that the ”high” value was greater than the ”low” value for each trading day.

  The verification process was successfully completed, and it was confirmed that all stock price data followed the expected pattern, with every ”high” value being greater than the corresponding ”low” value. This ensures that the dataset is consistent in terms of stock price fluctuations.

Thanks to the successful outcomes of the two data quality metrics, **Temporal Completeness** and **Data Consistency**, and considering that the data source (Alpha Vantage) is well-regarded for their reliability, we are confident that the collected data is of high quality. These results allow us to proceed with the analysis, confident that the data foundation is solid and free of significant issues.

4.2 Data Quality on Insider trading

- **Data transaction and publication consistency** In this phase, we checked the consistency between the transaction date (`Date_transactions`) and the publication date (`Date_publication`). The rule imposed by the SEC should implied that the transaction date must always precede the publication date. This ensures that the insider trading data is consistent and correctly ordered.

  The results of the validation are as follows:

  - **Total checks performed**: 18,914
  - **Valid checks**: 18,246
  - **Failed checks**: 668
  - **Checks with 'NA' dates**: 2

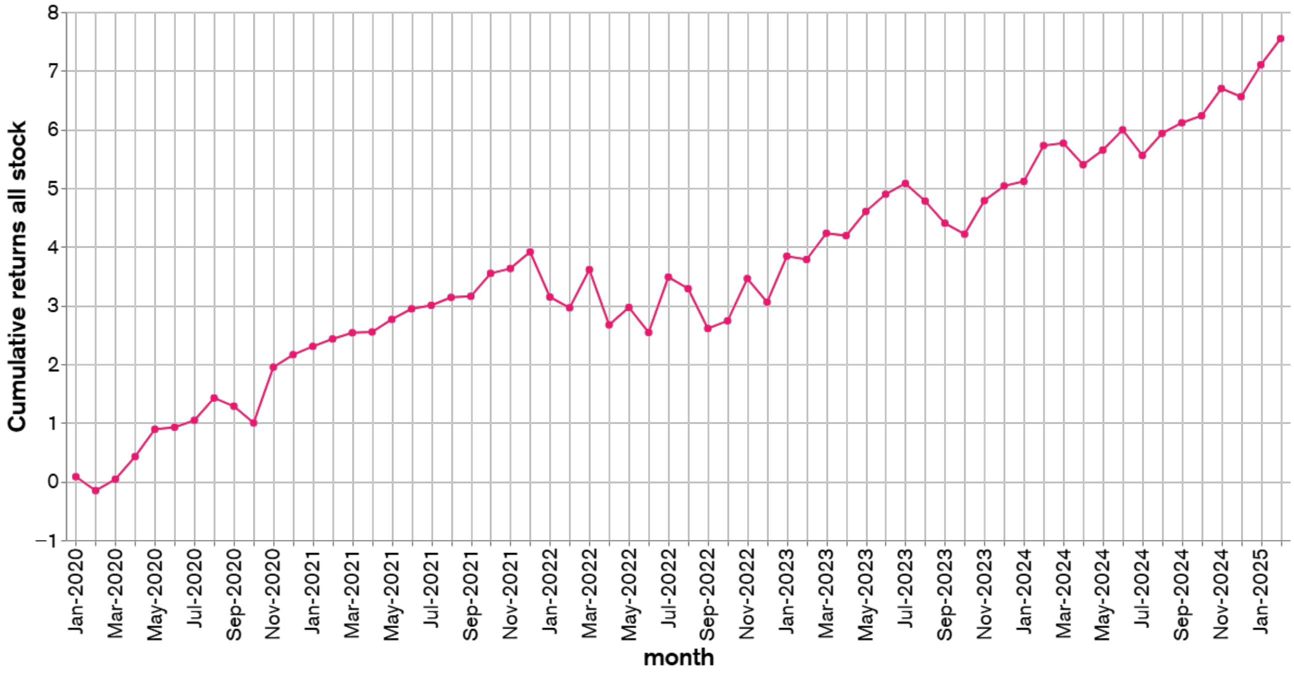
  This shows that 3.53% of the checks failed, meaning there were inconsistencies where the transaction date was not before the publication date. However, the number of valid checks is high, indicating that the majority of the data is consistent with the expected order of transaction and publication dates.

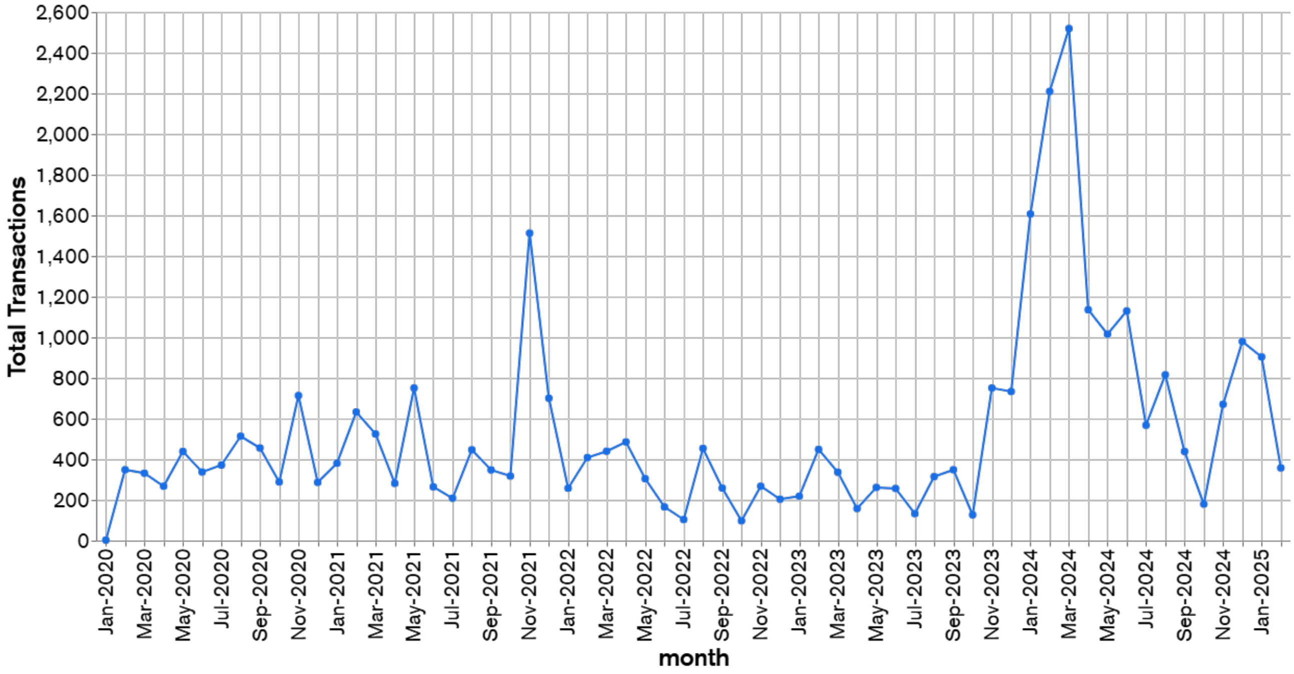
4.3 Data Quality on Integrated Data

- **Completeness** This metric checked the completeness of the data by counting how many field have null values. It turned out that 37972 values are null on 1282676 values evaluated, it is only **0. 02%**.
- **Price Range Consistency** This metric checked whether the purchase price of insider traders ("transactionPricePerShare") falls within the high/low range for the corresponding stock in that day. This metric helped refine our initial filters selecting only the relevant transaction codes for our analysis, while excluding others that were not subject to this constraint. Moreover, during the process, this check helped us to spot a bug in the scraper as it was wrongly including some data from different person or company.

5 Visualizations

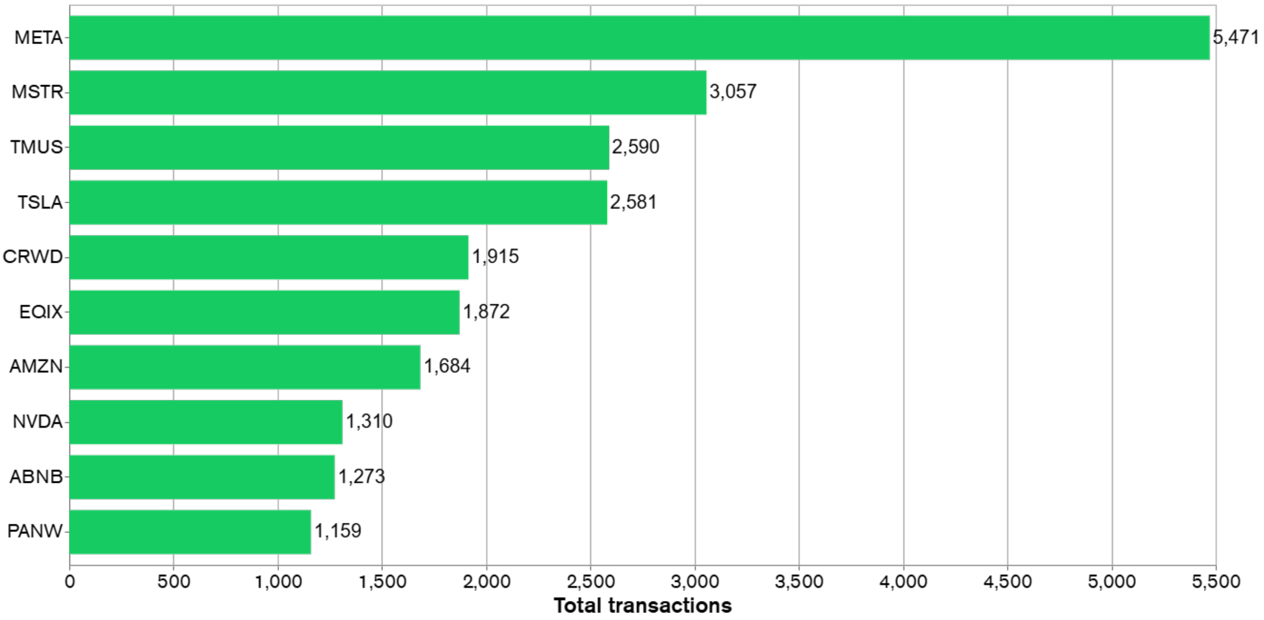
In the end we can use the integrated dataset to discover the hidden structures of the data and discover eventual relationships between the insider transactions and the stock price. (All the code of the queries is available in the file "data viz query"). First of all, giving our filters, the insider transaction data are composed by **95% sell transactions**. This actually could be logical as the insider typically receives the stock as part of salary and eventually they decide to sell some shares on the market, also as a way of reducing the exposition on one single company. Keeping this in mind, we can inspect in the following graph if the transactions (almost all sales) have an impact on the prices. They can have this impact because of the demand and supply balance but also potentially because they may know in advance that the market is reaching a peak.





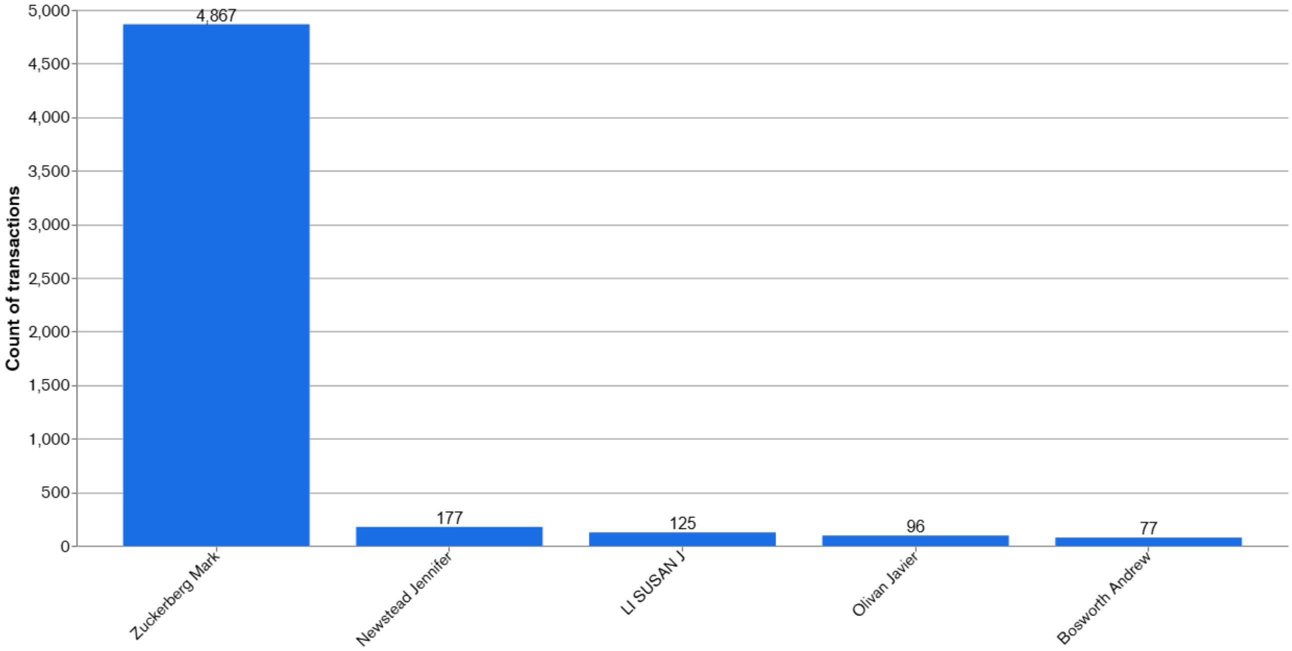
We can see that, for example, in the peak of transaction on the November 2021 the overall performance of the stock in our analysis decrease also incurring in a slight drawdown. We can also see that on the other major peak of transaction in March 2024 the price slow down from an uptrend. This is an aggregated result across all the companies for restricted time window but this integrated data allows us to explore specific aspects about a single stock and differences across them.

For example we can ask how many transactions in the last 5 years have each company, these are the top 10 companies:

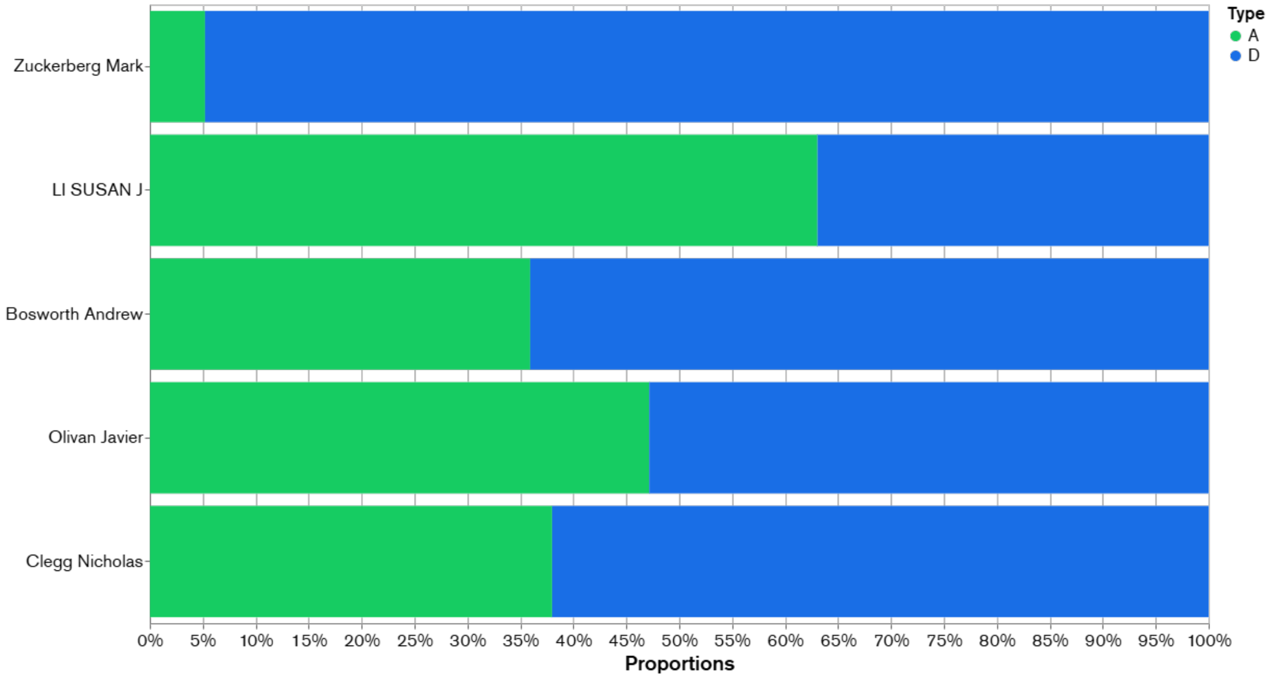


META clearly has an impressive number of transactions respect to other company. Maybe it is due to the fact that there are many insiders that need to compile form-4. But if we count the distinct entity we obtain 13 people that is slightly below the mean of other stock (that is 15 people).

Now we can inspect who among those people carried out the most number of transactions, here are the top 5:



We can compute if these different people executed most buy or sell as insider in proportions:



We can see that the mostly transaction are sales for the CEO of the company, while Li Susan J, that is, the Chief Financial Officer, in proportion bought more. This could indicate that she is confident about the future of the company, also considering the fact that in general buy transactions are quite rare in this data.

6 Conclusions

We successfully built a well-structured dataset that efficiently integrates two crucial aspects of the stock market: stock performance and insider trading activities. By combining these two dimensions, we created a comprehensive resource that enables a more in-depth analysis of how insider transactions relate to stock price movements over time.

From a broad perspective, our analysis suggests that insider trading may have a measurable impact on stock performance. In particular, we observed that certain periods exhibit a correlation between insider transactions and significant stock price fluctuations. This suggests that insider trading could be a leading indicator of market movements, potentially offering valuable insights for investors and analysts.

The fine-grained structure of our dataset allows for an in-depth examination of individual stocks, providing the ability to analyze how different types of transactions—such as acquisitions and dispositions—are executed by key figures within a company. This granularity is essential for distinguishing patterns in insider behavior and understanding their potential influence on market trends.

Future research could build on these findings by considering additional factors, such as market conditions, industry trends, or major economic events that might influence insider trading. Expanding the dataset to include more stocks or a longer time period could also help validate the results, maybe by doing some statistical tests or machine learning models. This could provide a clearer picture of insider trading patterns and their potential impact on financial markets.