

# Final Reflection: U.S. Congress Stock Trade Analysis

Yinka Vaughan  
QAC 420: Data for Good

April 2026

## Project Summary

This project set out to test whether members of Congress, using only publicly disclosed financial filings, outperform the broader market. By scraping and standardizing disclosures from both the House and Senate, merging transaction records, and benchmarking those trades against the S&P 500, I built a reproducible analytic pipeline that transforms fragmented government disclosures into clear evidence about financial performance and information advantage.

## What I Learned

The biggest takeaway is how much the project was a data engineering challenge as much as a statistical one. The disclosures are public, but they are not easy to work with. I had to combine web scraping, PDF parsing, and browser automation to collect complete data sets. Cleaning the resulting tables required careful regex work and column normalization so that trades from different chambers, filing formats, and document styles could be merged into a single analysis-ready table.

On the analysis side, the project helped me translate financial concepts into measurable metrics. Calculating 30-, 90-, and 180-day returns against the S&P 500 allowed me to define Jensen's Alpha as the difference between congressional return and expected market return.

That made the core question concrete: are these trades merely riding the market, or are they generating excess return beyond what the public could expect from the same time frame?

## Key Findings

The final analysis found a positive excess return for tracked congressional purchases. In the dataset I built, the average 90-day return on those purchases was meaningfully higher than the 90-day market benchmark, producing an average 90-day alpha of about 2.58%. A 1-sample t-test comparing that alpha distribution to a market-neutral expectation of zero returned a statistically significant result at the  $p < 0.05$  level, suggesting this outperformance is unlikely to be pure chance.

That result does not mean every member of Congress beats the market. In fact, most legislators did not show consistent outperformance; the story is strongest in the tails. A small subset of trades and traders drove the most compelling alpha, which is why the presentation focuses on the anomalies and the people behind them.

## Ethical Reflection

A major ethical theme of this project is accountability through public data. I chose to work only with disclosures that are legally required and already available to citizens. The scraping did not bypass any protected systems or violate access controls; it automated the reading and structuring of data that is meant to be public. This made the project an exercise in unlocking transparency rather than in exploiting privacy.

At the same time, I remain conscious that the data is imperfect. Not every trade is captured cleanly, and filing delays can make transaction timing ambiguous. Those limitations must be part of the final message: the analysis is evidence of a pattern, not a definitive proof of misconduct.

## Technical and Analytical Challenges

- Data collection: Parsing inconsistent PDF disclosure formats and bypassing visible-only table pagination on the House site required browser automation with Playwright and PDF table extraction with `pdfplumber`.
- Data quality: Normalizing asset names, handling missing ticker matches, and reconciling duplicate rows were essential so the benchmark calculations would be valid.
- Benchmarking: Building a reliable S&P 500 comparison meant pulling market prices for the right date windows and computing returns for each transaction date, then calculating alpha as excess return.
- Statistical interpretation: Presenting the t-test and  $p$ -value in a way that an audience can understand without technical background was an important communication challenge.

## What I Would Do Next

For future work, I would expand the data coverage to more transaction types and incorporate committee assignment as a control variable. That would allow a stronger test of whether certain decisions are linked to access to non-public legislative information. I would also refine the timeline assumptions for filing delay to reduce noise in the 30-, 90-, and 180-day return windows.

## Final Thoughts

This project ended up being more than a question about market timing. It became a study of how public accountability can be turned into data. The dataset and code make an otherwise opaque set of filings accessible and comparable, and the analysis gives tangible meaning to

the idea of congressional financial advantage. As a final takeaway, the strongest contribution here is the workflow itself: a repeatable pipeline that can be updated as new disclosures are released and reused for future accountability research.