**Bai**, Jeffry; **Blumrosen**, William; **Nyamalor**, Kelvin

IST 718 Final Project

Financial Market Analysis

March 2024

# 

# **A Data Scientist's Guide to the Financial Market**

## Abstract

The financial market is ripe with vast amounts of structured and unstructured data, a dream for data scientists to hone their technical and analytics skills. In the ever evolving field of financial market analysis, the incorporation of unconventional data sources such as politician trade activities and news sentiment has garnered increasing interest among researchers and practitioners alike. This project embarks on an empirical exploration to assess the potential impact of these two signals on stock market movements. Leveraging a comprehensive dataset that encapsulates politician trading activities alongside sentiment analysis of media coverage, we aim to explain the extent to which these factors influence stock prices over a 30-day trading period.

Contrary to the hypothesized significant impact, our findings reveal a modest edge of only 2% attributed to politician trade signals, underscoring a nuanced interaction between such trades and stock performance for a short term strategy. Additionally, the predominance of neutral news sentiment further complicates the observable influence on market dynamics, suggesting that the predictive power of media sentiment on stock movements is less pronounced than anticipated. This paper delves into the methodologies employed for data collection and analysis, discusses the implications of our findings within the broader context of financial market prediction, and outlines avenues for future research to enhance the understanding and application of these novel data sources in market analysis strategies.

## SPECIFICATION

This analysis endeavors to investigate the influence of politician trading activities and news sentiment on stock market movements, setting forth a nuanced inquiry into how these unconventional data sources might serve as predictors for financial markets. The primary hypothesis posits that both politician trade signals and news sentiment have a discernible impact on stock prices, potentially offering a strategic edge to investors and analysts. To rigorously test this hypothesis, the study utilizes a dataset comprising detailed records of politician trading behaviors over a corresponding 30-day trading period from the disclosed trade date. The problem centers on determining the precise magnitude and direction of these effects, if any, on stock market dynamics.

The datasets employed in this analysis, the study meticulously compiles and leverages diverse sources to ensure a robust examination of the posited hypotheses. Politician trading data is collected through web scraping techniques from CapitolTrades.com. To assess news sentiment, media extracts from Alpaca.Market with the application of sentiment analysis tools. Additionally, stock market information is sourced from yfinance. This unique amalgamation of datasets from CapitolTrades.com, Alpaca.Market, and yfinance forms the bedrock of our analysis, enabling a multifaceted investigation into the interplay between politician trade signals, news sentiment, and stock market performance.

## ANALYSIS AND MODELS

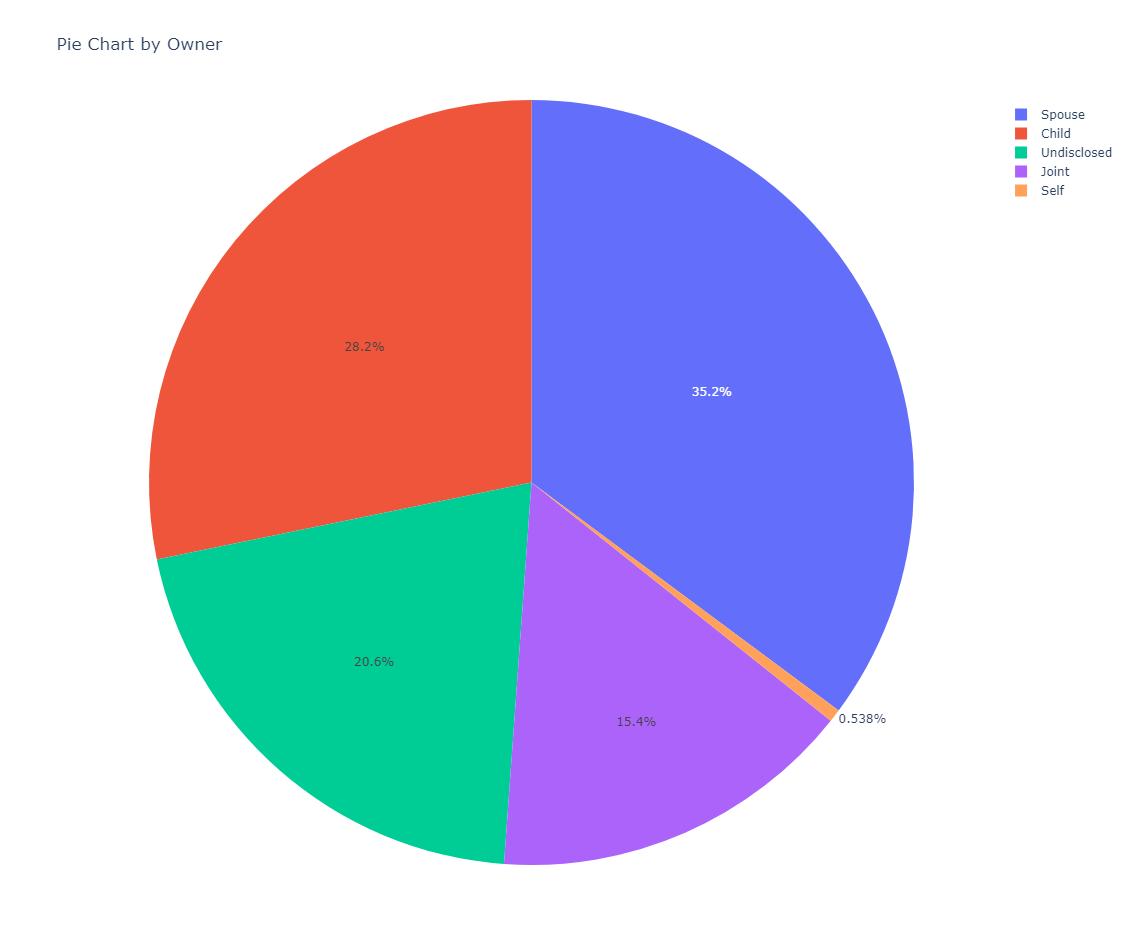
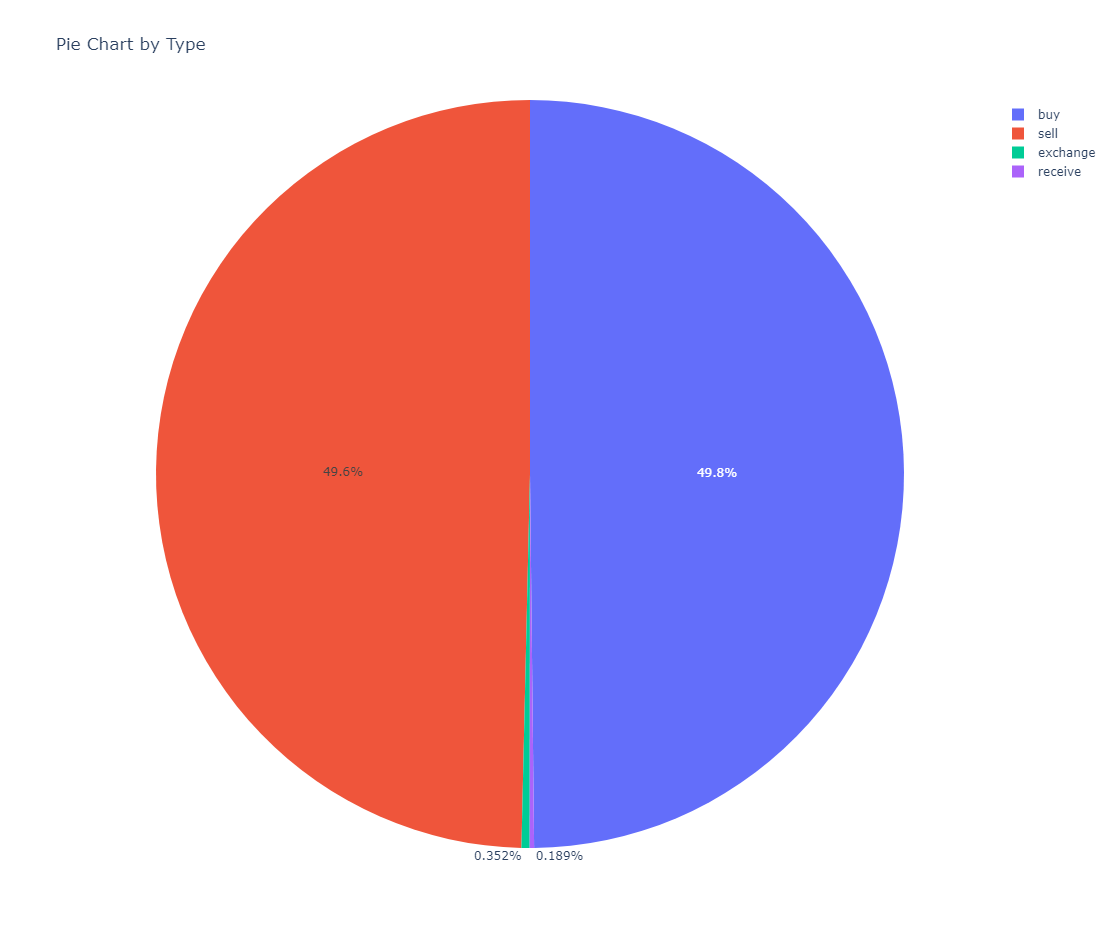
### About the Data

#### Politician Trades

The table scrapped from CapitolTrades.com consists of 34,425 entries and consists of 10 columns. The attributes includes the name of the politician involved in the trade, the traded issuer, publication date of the trade information, the trade date, the filing date after the trade, the owner of the trade, the type of trade, the size of the trade, the price at which the trade was executed, and the asset type involved in the trade. There are a total of 1,843 unique stocks being traded. Features of this dataset represented in the figures below.

| **Attributes** | **Type** | **Description** | **Label** |
| --- | --- | --- | --- |
| **Politician** | *Nominal* | Raw data that is the concatenated values of politician name, affiliate congressional branch, affiliate political party, and represented state. | **167** unique values |
| **Traded Issuer** | *Nominal* | Raw data that is the concatenated values of company name and ticker symbol if available. | **1843** unique values |
| **Published** | *Date* | Date that the financial disclosure is published. | **02/03/2021** to **02/02/2024** |
| **Traded** | *Date* | Date that the trade is made. | **02/02/2021** to **01/29/2024** |
| **Filed after** | *Integer* | Number of days between trade date and disclosure date. | **0** to **911** days |
| **Owner** | *Nominal* | Owner of the trade and their relationship to the politician in question. | Self, Spouse, Child,  Joint, Undisclosed |
| **Type** | *Nominal* | Buy or sell. | Buy / Sell |
| **Size** | *Nominal* | Size of the trade, bucketed into discrete values. | **9** bins |
| **Price** | *Float* | Price of the asset at the time of trade. | **29748** non-null values |
| **Asset Type** | *Nominal* | Type of the Asset that is being traded. | Stocks and Crypto |
| **State** | *Nominal* | State that politician represents. | **45** unique values |
| **Congress** | *Nominal* | Congressional branch that the politician sits on. | House,  Senate |
| **Party** | *Nominal* | Political party that the politician is affiliated with. | Democrat,  Republican,  Other |
| **Politician Name** | *Nominal* | Name of the politician |  |
| **Ticker** | *Nominal* | Ticker symbol of the trade in question. | **1821** unique values |

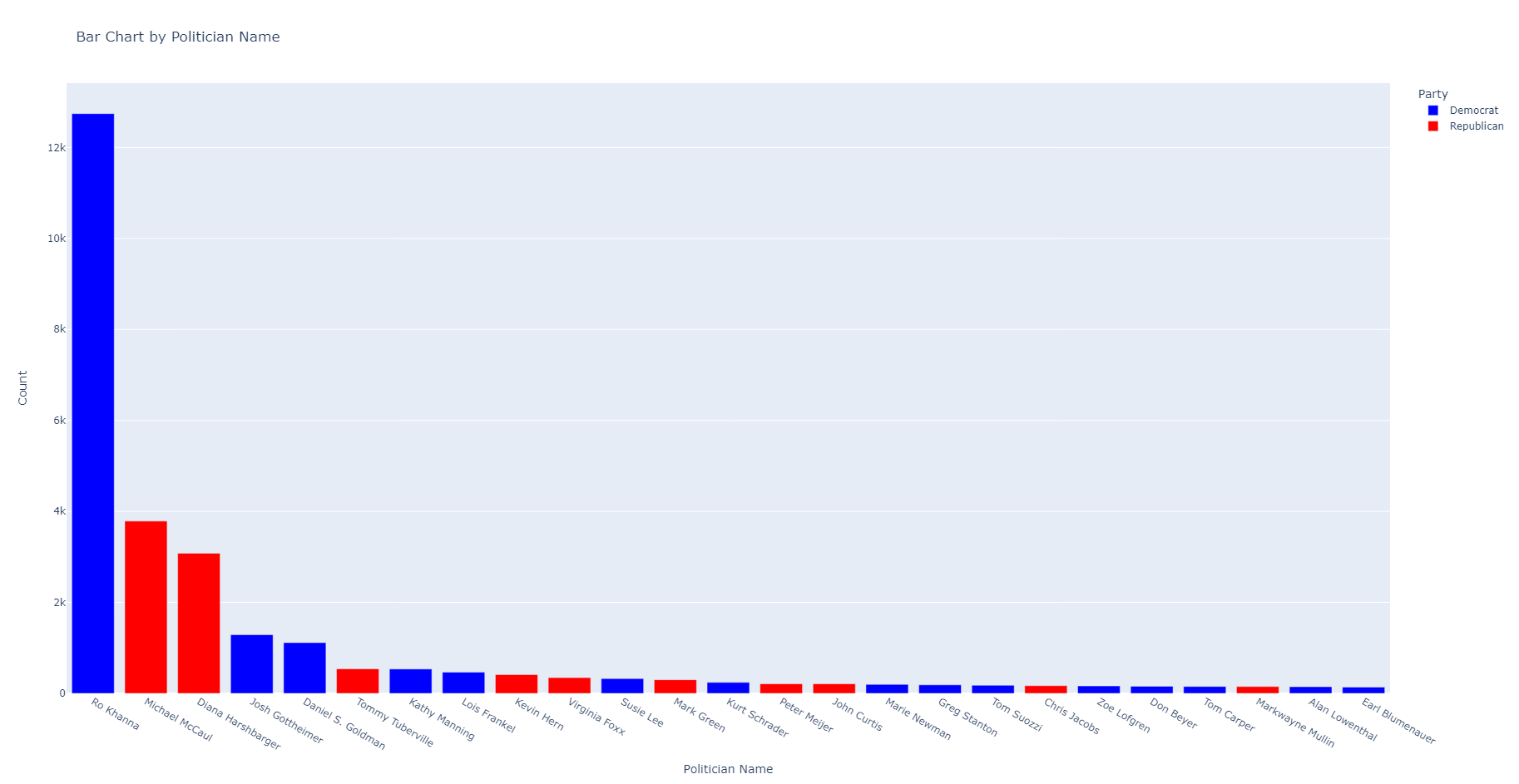
*Fig 1. Table of the final cleaned data from CapitolTrades.com*



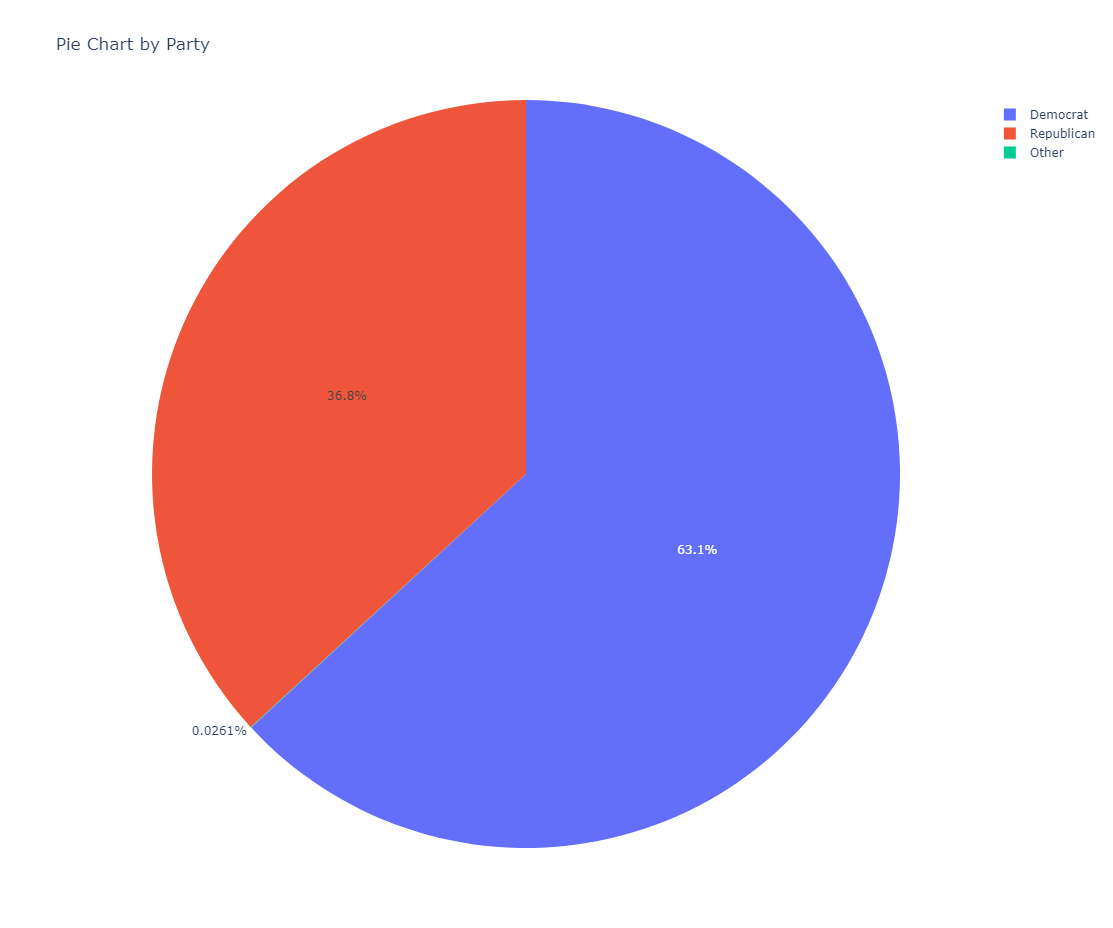
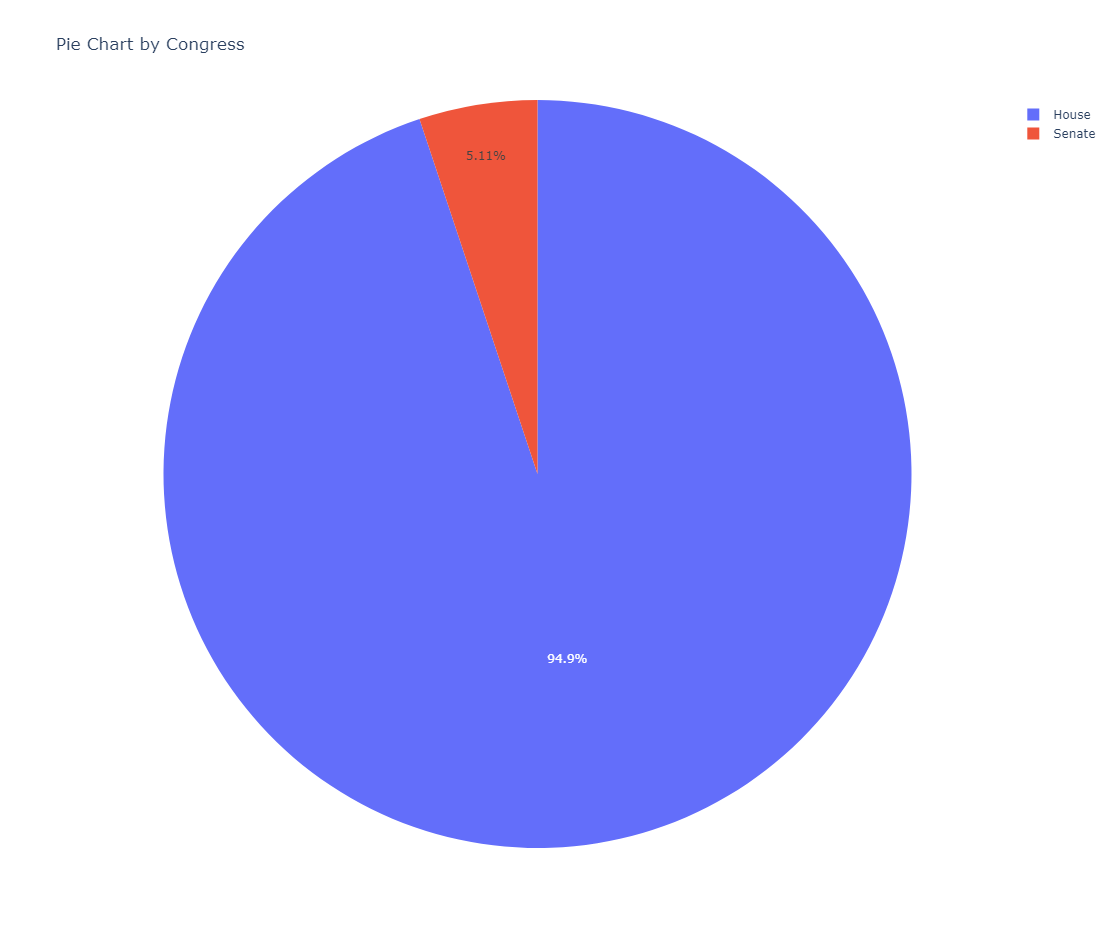
*Fig 2. Pie chart of trade type and by owner respectively.*

*Left: Trade type is about half buy and half sell, with negligible amount exchanged or received.*

*Right: Ownership is mostly by Spouse, then Child, then Undisclosed, then Joint, followed by a small amount by Self*



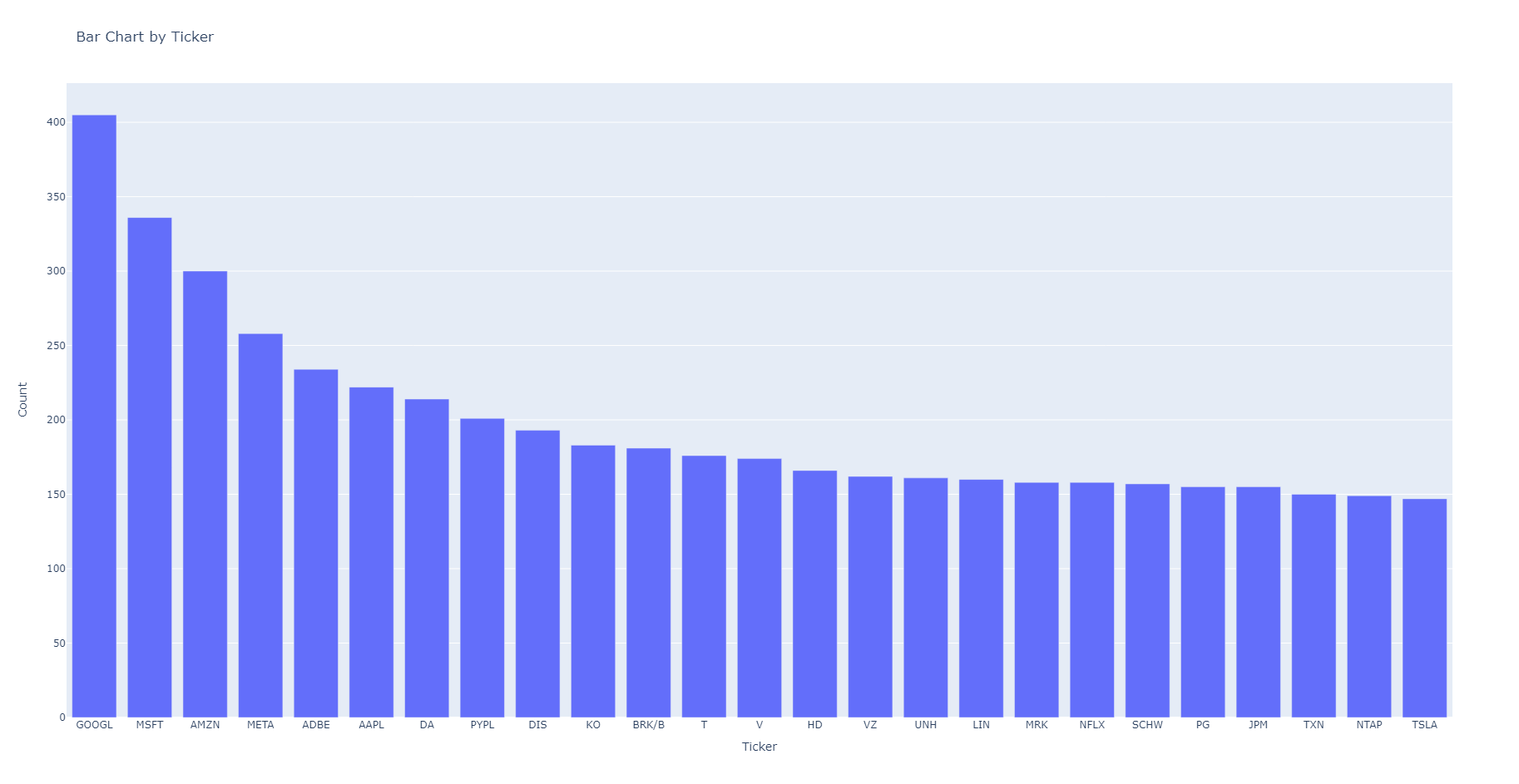
*Fig 3. Bar graph that shows distribution of trade by politician, color coded by political affiliation. Representative Ro Khanna makes up about a third of the total trade volume examined, a quality to be kept in mind for the rest of this analysis.*



*Fig 4. Pie chart of trade by congressional branch and political affiliation respectively.*

*Left: Trade by congressional branch. The House makes up the majority of the trade since there are more Representatives than Senators.*

*Right: Trade by political party. Democrats make up about two-thirds of the trade; however if Rep Khanna is removed, the split would be more evenly split between Democrats and Republicans.*



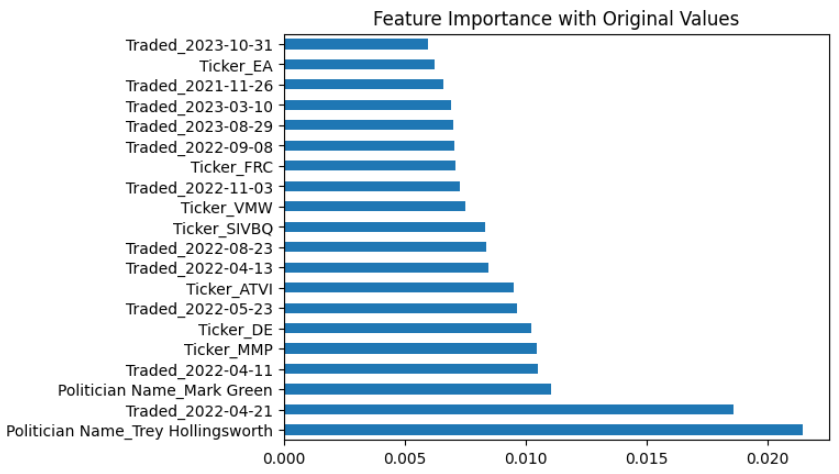
*Fig 5. Bar chart that shows the trade by ticker symbol. Notably the stocks that had the most growth in the past few years had the most volume of trade.*

Additional data was brought in from yfinance to compliment the politician trade data. Prices from the traded date, published date, and 30 days from traded date were added to the final dataframe. The differences between the published and traded price as well as 30 days from traded to traded price were computed and added for analysis. To compute expected return, these delta were multiplied by -1 if the trade type was a sell, so as to represent loss or unrealized gains if negative, and gains or loss saved if positive. These numbers are then standardized by converting it to a percentage to facilitate comparison.

Alpaca.Market was used to bring in news data for each stock and give a sentiment analysis of the news related to a stock on the spread of 3 days. I.e. If Google stock was traded on March 15th, 2022, the Alpaca.Market API was set to pull data from March 14th, 15th, and 16th to predict whether the news was positive, negative, or neutral on that particular stock. The way the news was analyzed was through a HuggingFace model called Finbert, which is a pre-trained NLP model on measuring the sentiment analysis of financial related news articles. Combing these things together helped us predict the sentiment of news for each stock traded on their respective days.

## RESULTS

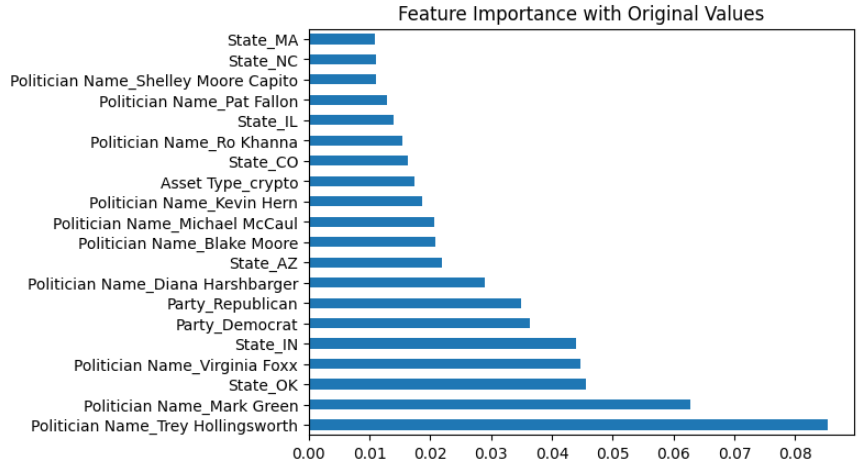
For this project, a boosted tree model was employed, an ensemble technique for its ability to improve prediction accuracy by combining multiple weak learners into a single strong learner. Boosted trees iteratively adjust to correct the errors of previous trees by placing more weight on the instances that were incorrectly predicted, thereby enhancing the model's ability to capture complex patterns and relationships within the data. Specifically, the model was trained using data encompassing politician trading activities and stock market information to predict stock movements. Boosted tree model was able to achieve an accuracy of 69% predicting buy or sell based on a 30 day trade window from the date of politician trade.



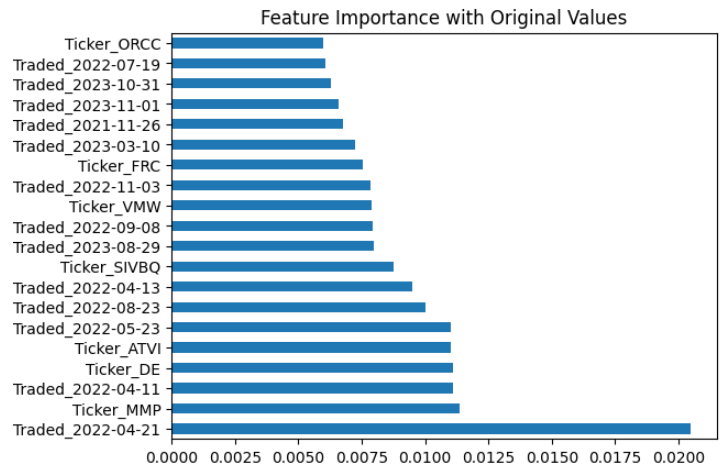
*Fig 6. Bar chart that shows the most important features for the boosted tree model with a 69% accuracy.*

In an effort to further refine our understanding of the model's performance and to identify the most influential features, a systematic process of attribute removal was applied. This involved selectively omitting certain attributes from the dataset and observing the resultant impact on the model's accuracy. First, the features related to stocks and dates were removed, leaving only features related to political affiliation. The accuracy dropped to 52%, indicating that purely trading on political affiliation is marginally better than a coin toss. While a 2% edge is still an advantage, it is not a strong support that political affiliation is an advantage for a 30-day short term strategy. On the other hand, after political affiliations were removed, leaving only features related to stocks and dates, the accuracy maintained at 69%. This is stronger evidence that the model’s accuracy was predicting whether particular stocks on particular time periods are signals for buy or sell. When size of trade was removed, the accuracy had a marginal decrease to 68%, suggesting looking at huge trades can potentially offer additional insights.

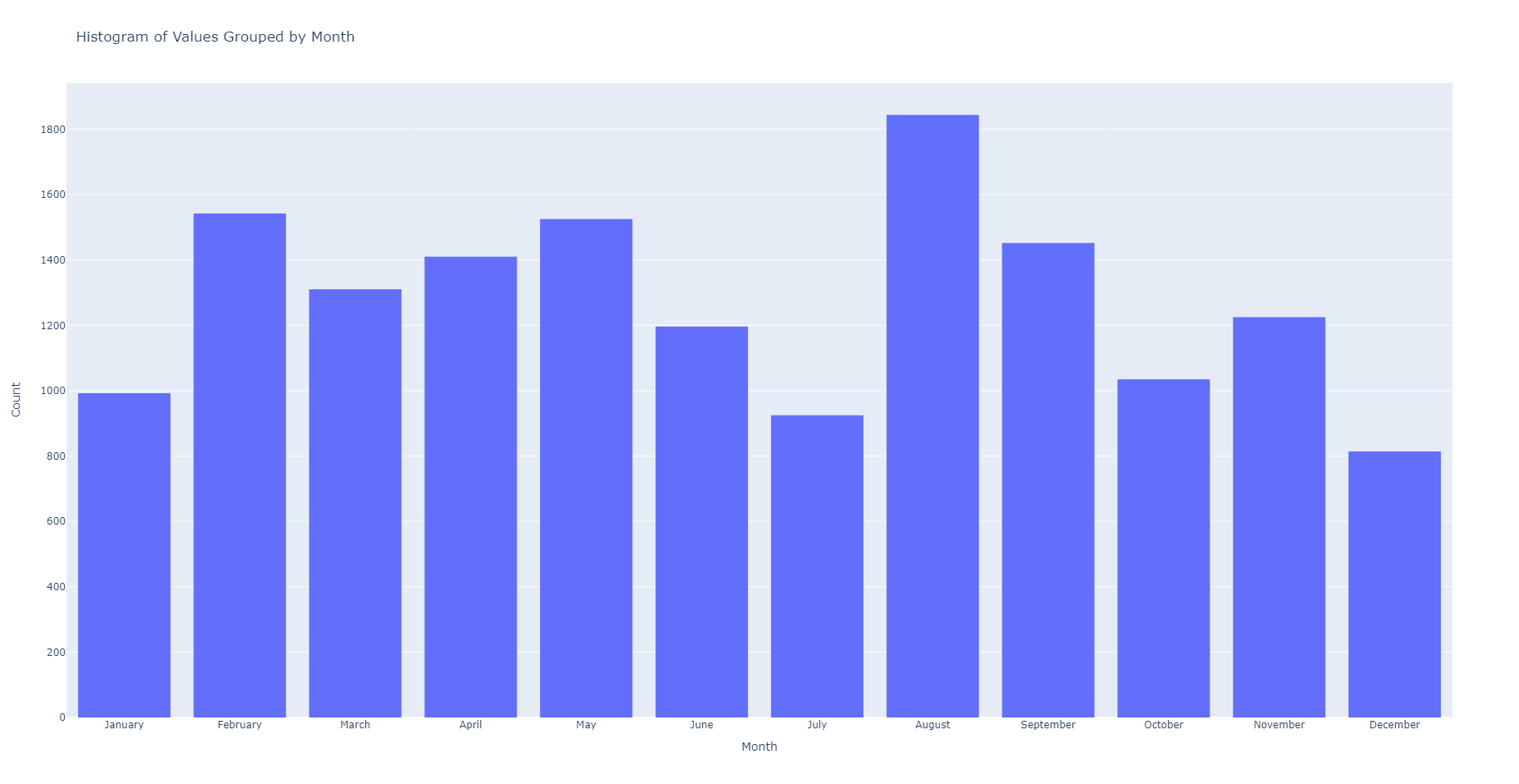
Since dates seem to have an impact on accuracy, this project delved into the traded dates to observe any patterns. A lot of activities occur in Q1 and Q2, which are the time periods for financial disclosures of most public companies in the prior fiscal year. A spike in trading activities in August, which is when congress is in recess. The spike in August is one of the few indications that the trading activities have certain political affiliation.



*Fig 7. Bar chart that shows the most important features for the boosted tree model with a 52% accuracy after all features about stocks and dates were removed, leaving only features related to political affiliation.*

**

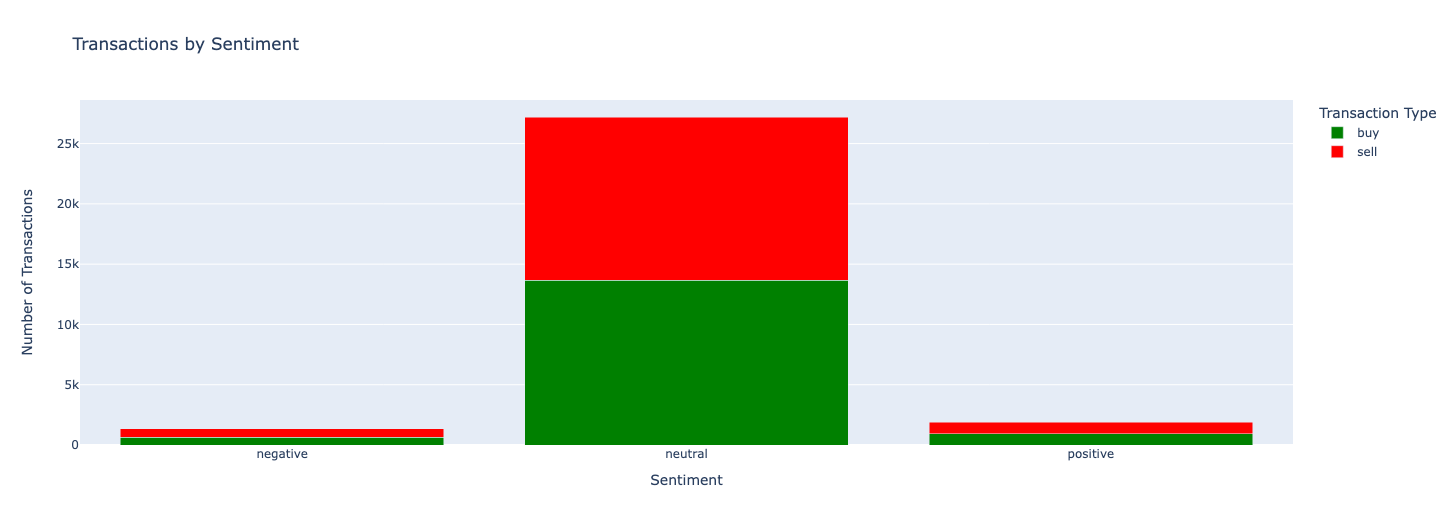
*Fig 8. Bar chart that shows the most important features for the boosted tree model with a 69% accuracy after all features about political affiliation were removed, leaving only features related to stocks and dates.*



*Fig 9. Bar chart of traded date by month.*

## *Sentiment Analysis*

After applying the finbert model to each and every stock trade in the dataset. Over 25,000 of the trades were classified as neutral and roughly 4,000 - 5,000 of them were positive or negative. Current news didn't seem to have much of an effect on a politician's decision to conduct a trade.



## CONCLUSION

The investigation into the predictive power of a boosted tree model, utilizing data on politician trading activities and stock market information to forecast stock movements, has yielded insightful conclusions about the dynamics influencing stock market predictions. Achieving an accuracy rate of 69% in predicting buy or sell signals within a 30-day window post-politician trade underscores the potential of leveraging complex patterns and relationships inherent in the data. However, the nuanced exploration into the impact of various attributes on model accuracy has further refined our understanding.

The systematic attribute removal process revealed that features related to political affiliation alone provide a marginal predictive advantage over random chance, suggesting that political affiliation, while interesting, does not significantly drive short-term stock movements. In contrast, maintaining attributes related to stock specifics and timing without political affiliations preserved the model's accuracy, highlighting the importance of these factors in predicting market behavior. The slight decrease in accuracy upon removing trade size data points to the potential value of analyzing the magnitude of trades as an additional layer of insight.

Moreover, the seasonal and periodic patterns observed in trading activities, particularly the spikes in Q1 and Q2 and during the congressional recess in August, hint at a complex interplay between political events and stock trading strategies. These patterns may offer a fertile ground for future research, particularly in understanding the temporal dynamics and their relation to financial disclosures and political cycles.

Lastly, news sentiment on particular stocks did not seem to affect the decision of a trade. It is very likely that politicians are not necessarily “day trading” stocks, rather buying them up for long term holds. If that is the case, then news sentiment not having an effect on the trade makes more sense.