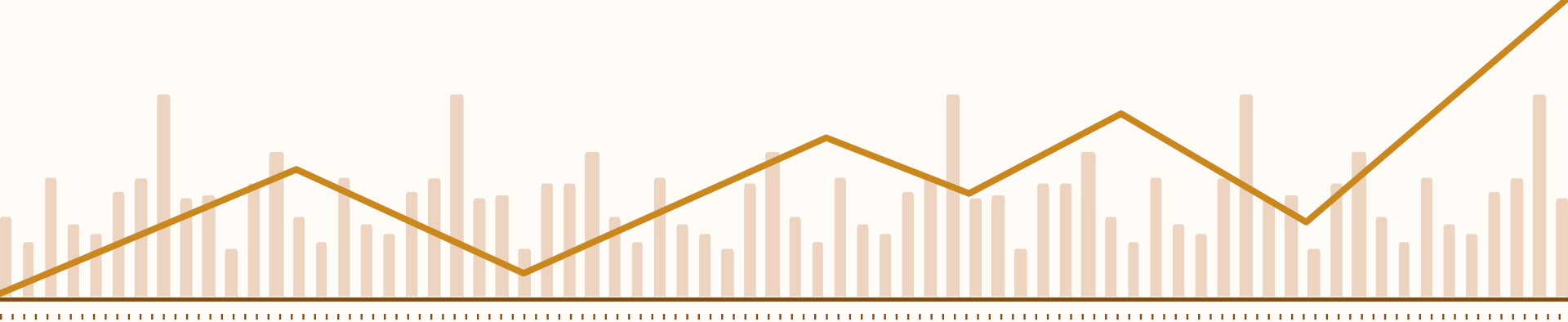


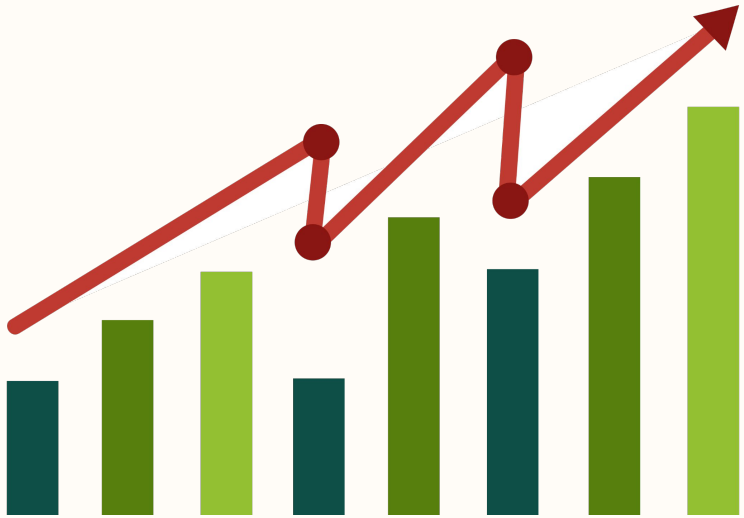
# Using a Hybrid LSTM-ARIMA Model to Forecast NDXT-100 Performance

Yuca Chen, Jacklyn Clauss, Skyler Lindsey, Dayanara Yanez

PHYS664 Midterm Presentation | March 20th, 2025

# Table of Contents

# Background



**Stock** (from Merriam-Webster): “part of the ownership of a company that can be bought by members of the public”

Average person’s interests in stock market has increased over the past few decades [1]
Machine learning techniques can help make more accurate predictions [2]
**Most people would like to make money!**

# Main Question

## How accurately can we predict the performance of NDXT-100?

In a particular time period?


# Data Sources

**NDXT-100: Main Dataset**

- NASDAQ-100 Technology Sector
- An equal weighted index comprised of securities that are considered to be technological

| | Data Type | Number of Entries | Number of Valid Entries |
|---|---|---|---|
| **Trade Date** | DateTime | 4803 | 4802 |
| **Index Value** | Float | 4803 | 4802 |
| **Net Change** | Float | 4803 | 4802 |
| **High Price** | Float | 4803 | 4802 |
| **Low Price** | Float | 4803 | 4802 |

# Data Sources (cont.)

**AAPL Historic Stock Data and GOOGL Historic Stock Data: Supplemental Dataset**

- Apple Stock Data
- Google Stock Data

| | Data Type | Number of Entries | Number of Valid Entries |
|---|---|---|---|
| **Date** | DateTime | 2517 | 2517 |
| **Close/Last Price** | Float | 2517 | 2517 |
| **Volume** | Integer | 2517 | 2517 |
| **Open Price** | Float | 2517 | 2517 |
| **High Price** | Float | 2517 | 2517 |
| **Low Price** | Float | 2517 | 2517 |

# Data Sources (cont.)

**CPI data - Supplemental Dataset**

- Consumer price index → proxy for inflation rate
- Available monthly
- Note that this data set is organized with the year as the row and the month as the column.
    - Because of this, the 2025 row is “missing” data for March onward, as that data is not yet available.

| | Data Type | Number of Entries | Number of Valid Entries |
|---|---|---|---|
| **CPI** | Float | 422 | 422 |

# Data Cleaning and Integration

Bring together all four separate datasets into one pandas DataFrame

1. Unify all datasets under the same dates [done]
2. Future standardization of data to gain relative behavior
    a. Divide stock price by CPI?
    b. Additional future data integration?

```
merged_all_info.describe()
```

| | CPI | IndexValue_NDXT | NetChange_NDXT | High_NDXT | Low_NDXT | Close/Last_GOOGL | Volume_GOOGL | Open_GOOGL | High_GOOGL | Low_GOOGL | Close/Last_AAPL | Volume_AAPL | Open_AAPL | High_AAPL | Low_AAPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 80.000000 | 80.000000 | 80.000000 | 80.000000 | 80.000000 | 80.000000 | 8.000000e+01 | 80.000000 | 80.000000 | 80.000000 | 80.000000 | 8.000000e+01 | 80.000000 | 80.000000 | 80.000000 |
| mean | 268.039275 | 5736.497621 | 9.292133 | 5790.580089 | 5675.568981 | 83.955922 | 3.695266e+07 | 97.739190 | 84.843744 | 83.043946 | 98.786505 | 1.225911e+08 | 97.739190 | 98.786505 | 96.694474 |
| std | 26.539729 | 2638.365152 | 109.248208 | 2670.481158 | 2608.219285 | 43.824001 | 1.473377e+07 | 64.850660 | 44.352052 | 43.340103 | 65.470142 | 7.056292e+07 | 64.850660 | 65.470142 | 64.104162 |
| min | 236.525000 | 2102.664689 | -424.604916 | 2137.083486 | 2090.910392 | 27.165000 | 1.600613e+07 | 23.872500 | 27.290500 | 26.988000 | 24.116300 | 3.528143e+07 | 23.872500 | 24.116300 | 23.832500 |
| 25% | 245.378000 | 3536.448809 | -27.421121 | 3537.052936 | 3513.920414 | 47.531625 | 2.698290e+07 | 38.038125 | 47.868500 | 47.447500 | 38.138125 | 7.120070e+07 | 38.038125 | 38.138125 | 37.816875 |
| 50% | 257.277000 | 5008.726820 | 10.142135 | 5066.038761 | 4969.873610 | 64.739250 | 3.302262e+07 | 66.973750 | 65.610750 | 64.234250 | 69.366250 | 1.064704e+08 | 66.973750 | 69.366250 | 66.876250 |
| 75% | 296.284750 | 8108.494297 | 50.478459 | 8150.127249 | 7988.130917 | 122.763375 | 4.089553e+07 | 155.470000 | 123.045250 | 121.803625 | 156.192500 | 1.521625e+08 | 155.470000 | 156.192500 | 152.922500 |
| max | 315.664000 | 10832.367099 | 274.786761 | 10839.929986 | 10662.366556 | 182.990000 | 1.090090e+08 | 229.520000 | 183.880000 | 181.300000 | 229.650000 | 4.473492e+08 | 229.520000 | 229.650000 | 223.740000 |

**Figure 1:** A snapshot of the pandas command “df.describe()” on the main dataframe, “merged_all_info”, that combines all the data listed in the table above under an index of datetime. The total count of data points, mean, standard deviation, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value for each data column in “merged_all_info” is printed.

# Data Visualizations

Both stocks exhibit a general **upward trajectory** despite periodic fluctuations.

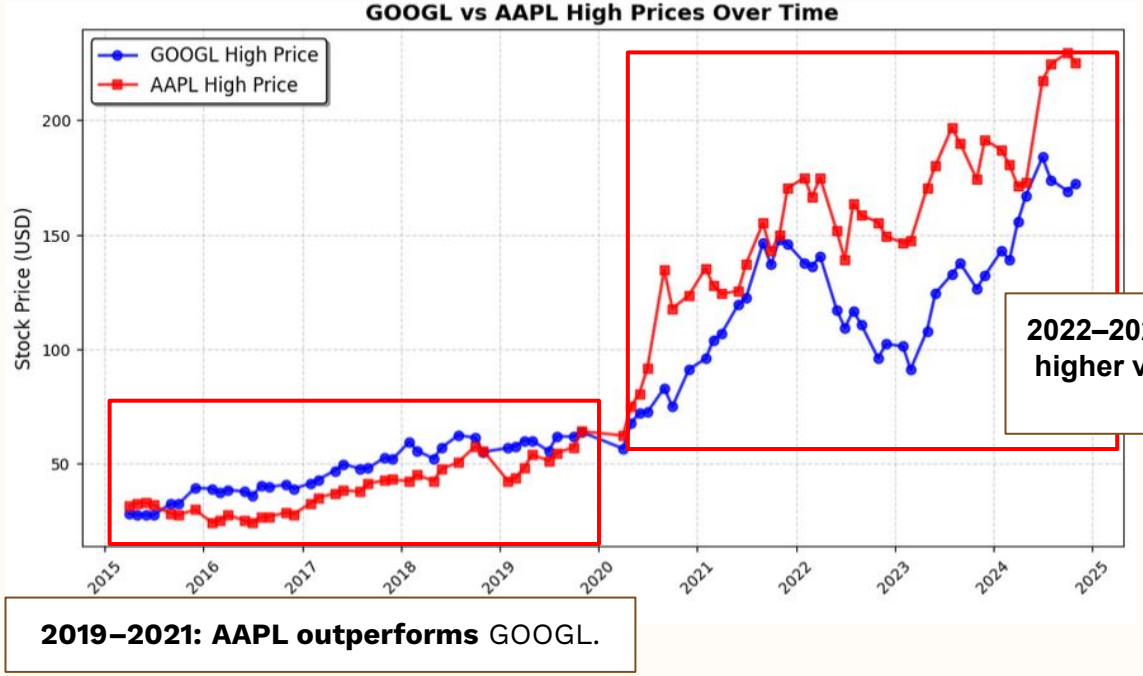



**Figure 1:** Trends are boxed off, highlighting fluctuations and performance differences over time.

# Data Visualizations (cont.)

Both stocks demonstrate **upwards and/or stable trading behavior.**

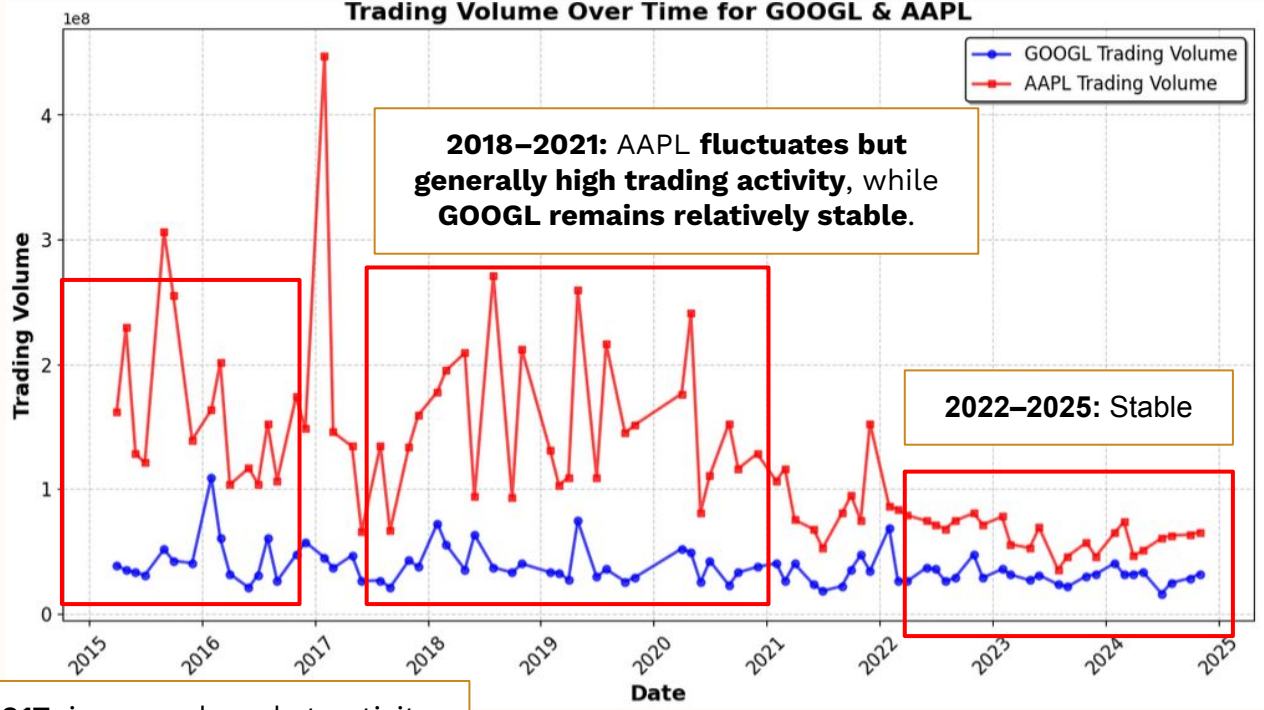
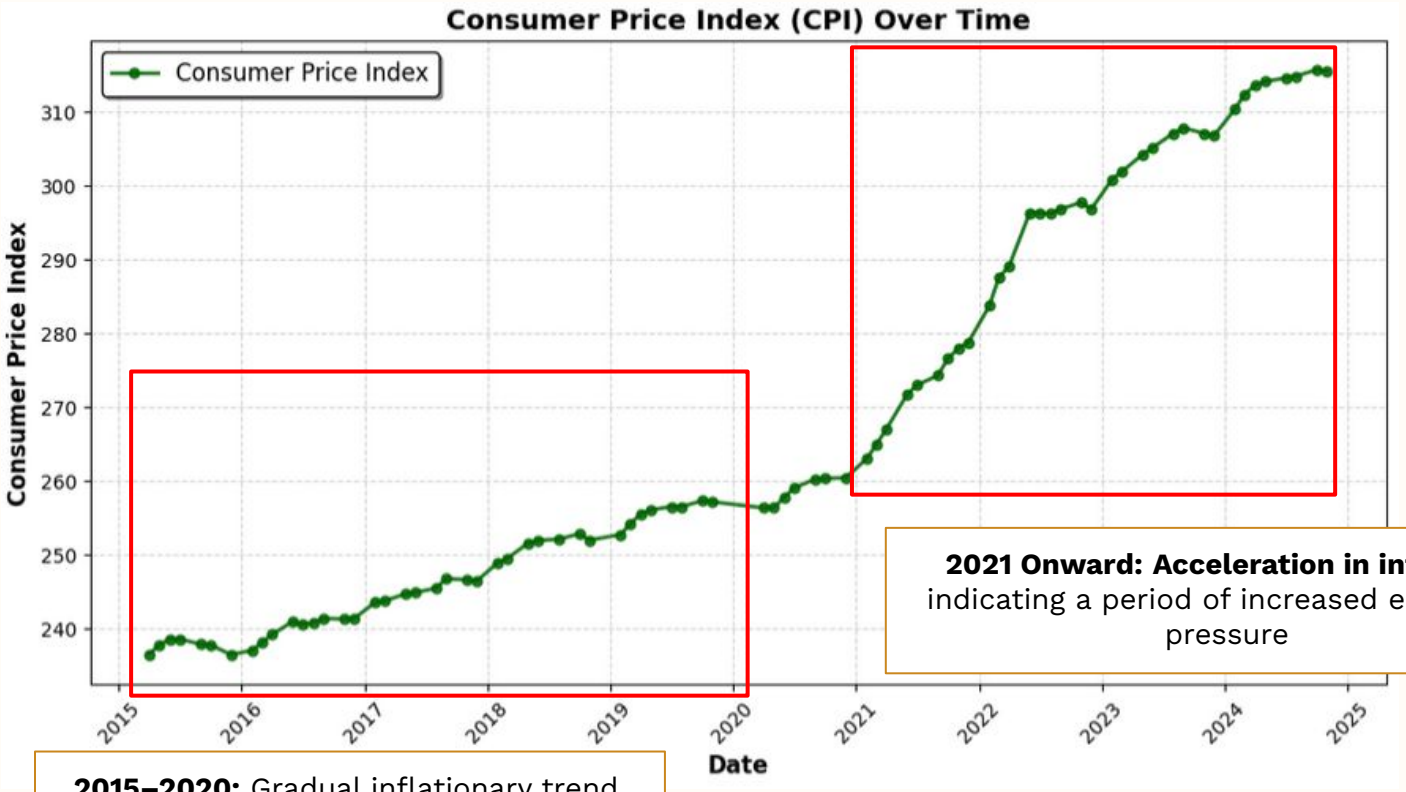



**Figure 2:** The differences in trading activity between AAPL and GOOGL are highlighted, providing insights into investor behavior and market liquidity over time.

# Data Visualizations (cont.)

The **CPI shows a steady increase**, reflecting rising costs of goods and services over time.




**2021 Onward: Acceleration in inflation**, indicating a period of increased economic pressure

**2015–2020:** Gradual inflationary trend.

**Figure 3:** CPI values on the y-axis and years on the x-axis, highlighting the impact of inflation over time

# Data Visualizations (cont.)

The **NDXT index** shows a strong **long-term upward trajectory.**

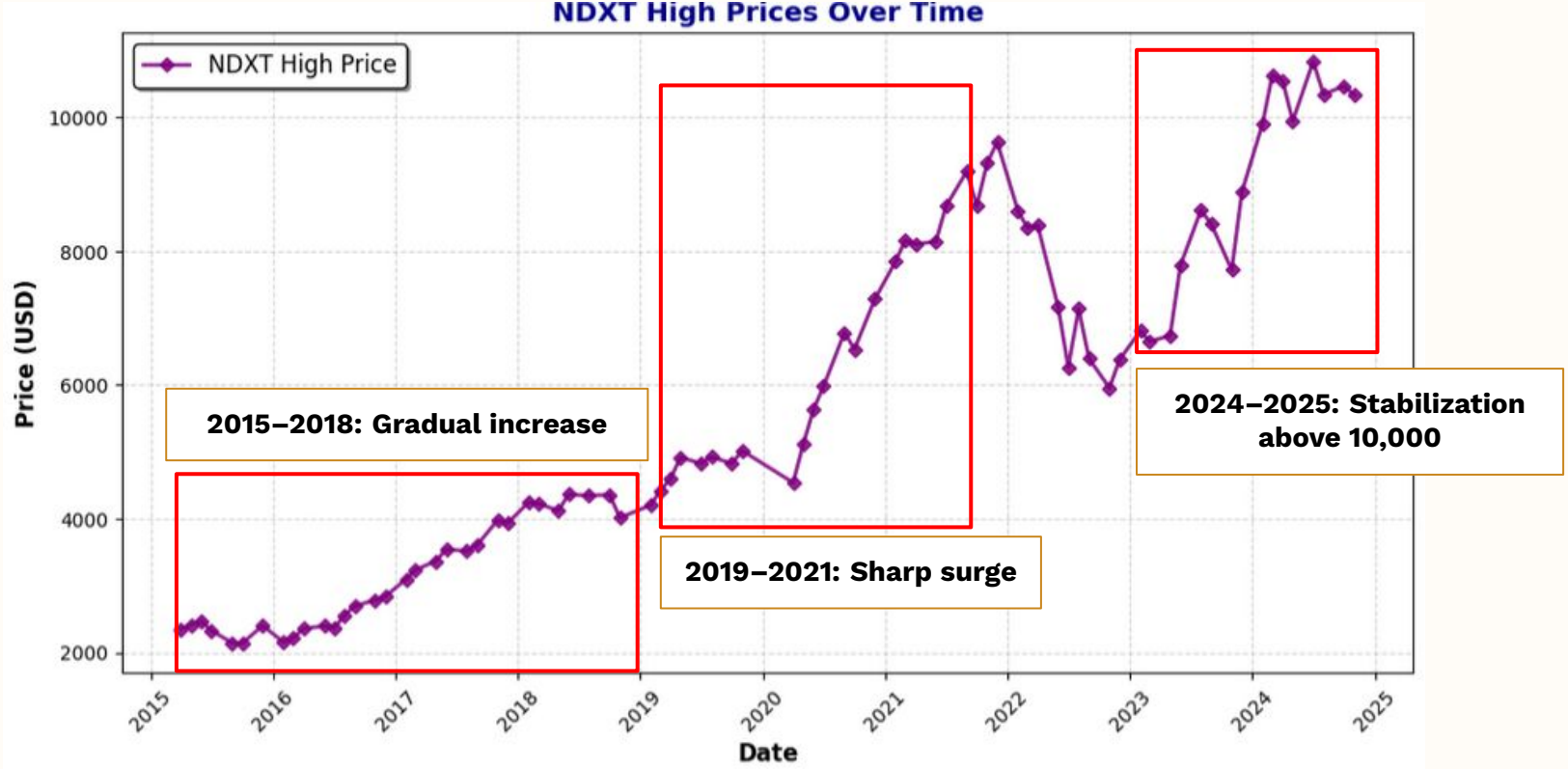
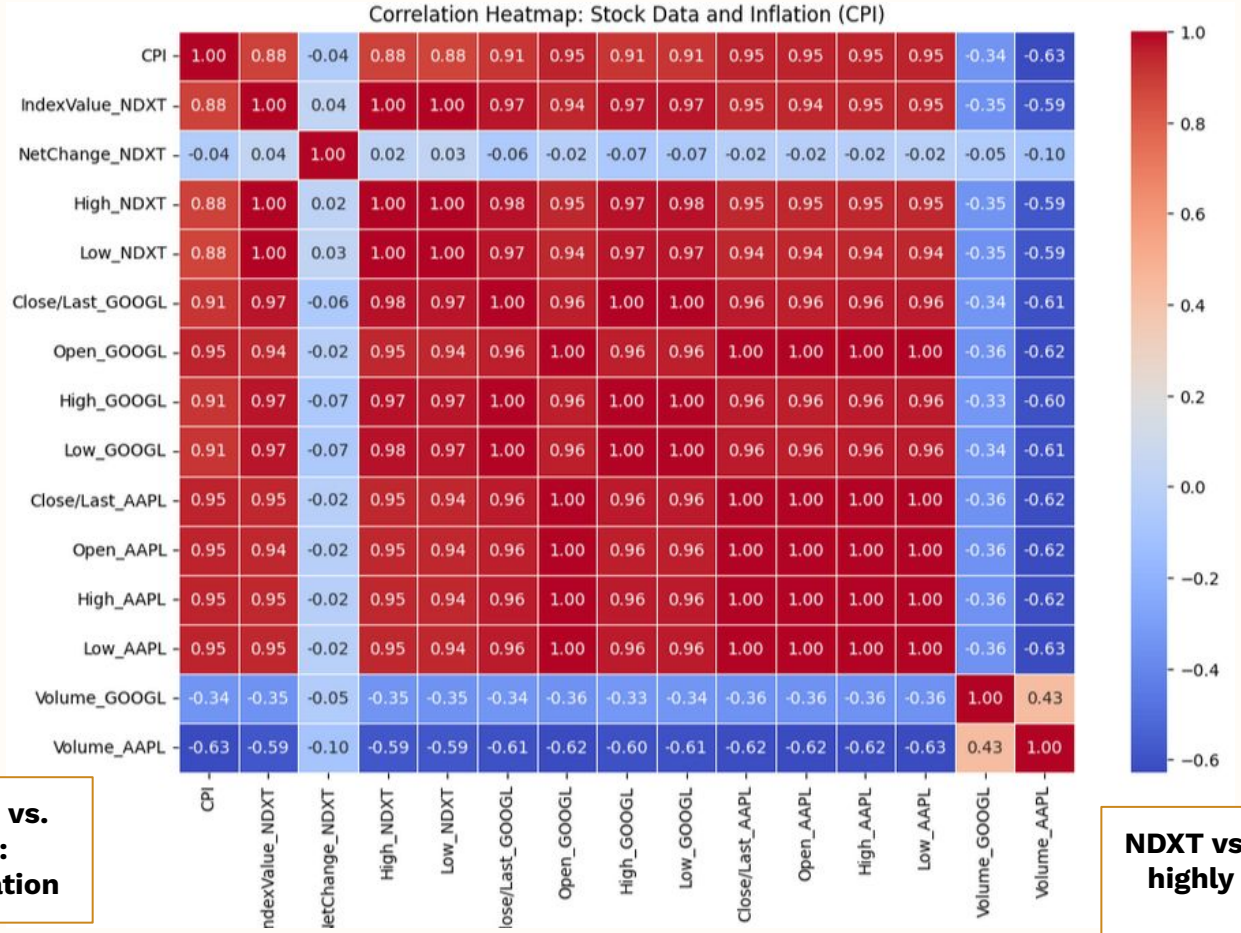



**Figure 3:** Trends in NDXT-100 are illustrated, emphasizing the long-term upward movement of the NASDAQ Tech Index.

# Data Visualizations (cont.)

**CPI vs. Stock Prices:** A **strong positive correlation (>0.90)**

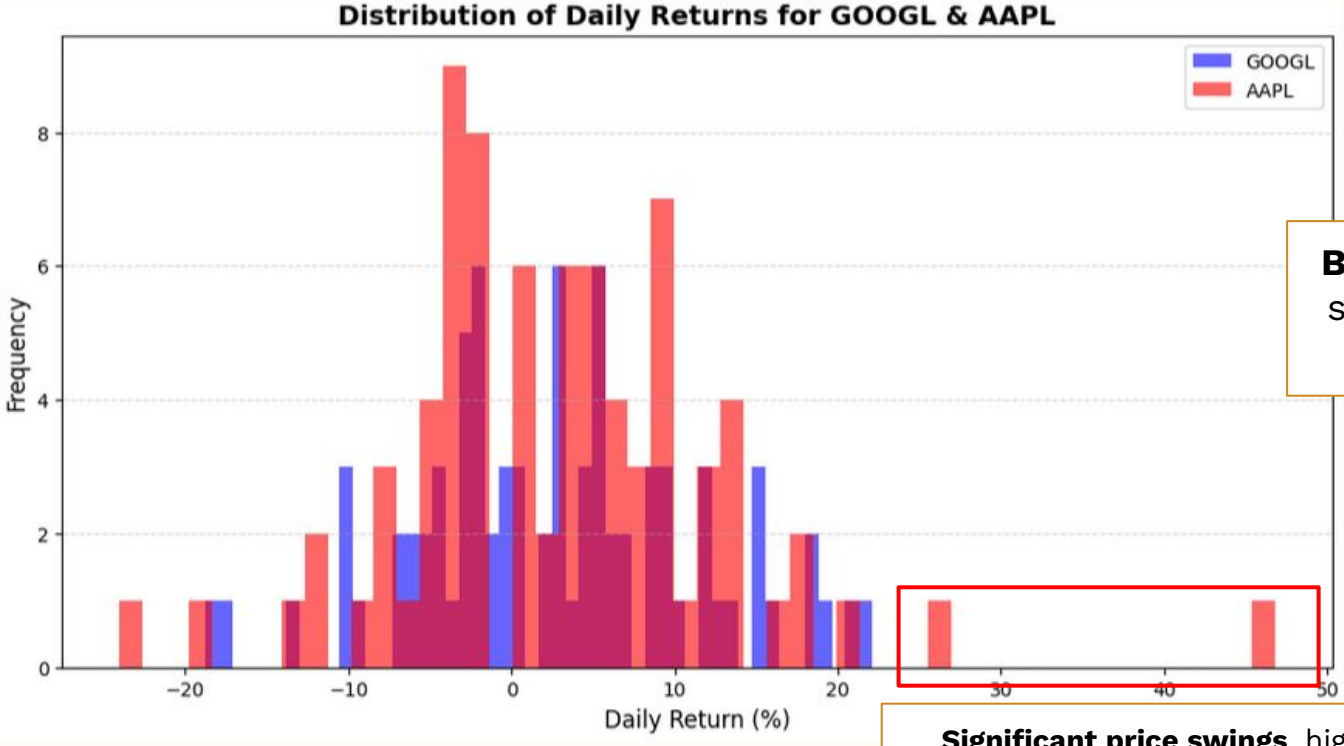



**Trading Volume vs. Stock Prices: Negative correlation**

**NDXT vs. Stock Prices:** NDXT is **highly correlated** with AAPL and GOOGL

**Figure 4:** The correlation matrix of all included datasets, highlights the interdependence between inflation, stock market movements, and trading behaviors.

# Data Visualizations (cont.)




**Bell-shaped distribution,** stability on most trading days

**Significant price swings**, highlighting periods of high market volatility.

**Figure 7:** histogram of daily returns (%) for AAPL (red) and GOOGL (blue) over a specified period.

Both stocks exhibit similar return patterns, though **AAPL shows slightly more extreme fluctuations** in both directions.

# State of the Art in the Literature

- Initial studies with regression, but lots of **recent** studies have used [3, 4]:
  - Artificial neural networks (ANN)
  - Support vector machines (SVM)
  - Long short term memory recurrent neural networks (LSTM RNN)
- **Range of different datasets** have also been called upon to feed into these algorithms
  - i.e. time-series data of available stock price information, macroeconomic information [2]

# Proposed Methodology

## Techniques Used:

### Long Short-Term Memory (LSTM)

A type of recurrent neural network that is effective in capturing long-term dependencies and non-linear trends in time series data

### Autoregressive Integrated Moving Average (ARIMA)

A statistical model designed for handling non-stationary time series data and capturing linear trends

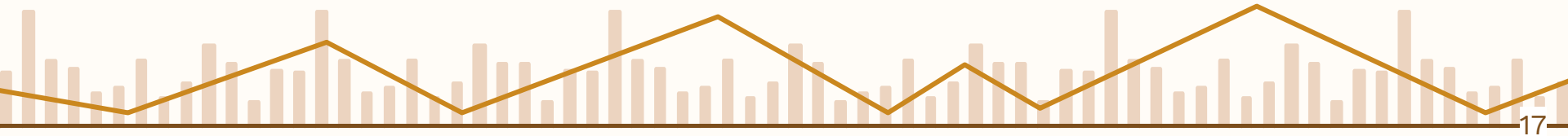
# Proposed Methodology (cont.)

## Why This Hybrid Approach?

LSTM excels in identifying complex patterns but may struggle with capturing overall linear trends

ARIMA provides error correction and enhances robustness against volatility

Combining both models allows us to improve **accuracy** in predicting stock market behavior by **capturing both local non-linear and global linear trends**



# Deliverable

1. A conclusion on whether these data sources are a significant predictor of the performance of the NDXT-100
2. A projection of our model into the future in order to attempt to predict the performance of the NDXT-100
    i. Determine a reasonable time frame for our prediction based on the performance of our model against test sets at different intervals of time
3. **A final prediction** of the performance of our main dataset in the best appropriate time interval.

# Future Work

1. Standardize data with respect to CPI
2. Create a Long-Short Term Memory neural network model to predict stock price
3. Integrate ARIMA into an LSTM model

# Project Points of Contact

| | |
|---|---|
| **Data Manager** | Jacklyn Clauss |
| **Communication Manager** | Yuca Chen |
| **Analysis Manager** | Yuca Chen, Skyler Lindsey |
| **Visualization Manager** | Dayanara Yanez |
| **Literature Manager** | Skyler Lindsey |

# References

[1] Badolia, Lokesh. How can i get started investing in the stock market. Educreation Publishing, 2016.

[2] Kumbure, Mahinda Mailagaha, Christoph Lohrmann, Pasi Luukka, and Jari Porras. "Machine learning techniques and data for stock market forecasting: A literature review." Expert Systems with Applications 197 (2022): 116659.

[3] Mintarya, Latrisha N., Jeta NM Halim, Callista Angie, Said Achmad, and Aditya Kurniawan. "Machine learning approaches in stock market prediction: A systematic literature review." Procedia Computer Science 216 (2023): 96-102.

[4] Liu, Yang. "Novel volatility forecasting using deep learning–long short term memory recurrent neural networks." Expert Systems with Applications 132 (2019): 99-109.

[5] Kashif, Kamil, and Robert Ślepaczuk. "Lstm-arima as a hybrid approach in algorithmic investment strategies." arXiv preprint arXiv:2406.18206 (2024).