

**University of Niagara Falls Canada**

**Summer 2025 Data Analytics Case Study 3 (DAMO-611-6)**

**Group 4 Members:**

**Akash Kumar (NF1015099)**

**Akash Joy (NF1003819)**

**Aenaben Jayeshbhai Parekh (NF1008546)**

**Nidhi Jaiswal (NF1006124)**

**Instructor – Omid Isfahanialamdari**

## **Phase 1: Problem Definition, Research Questions, and Hypotheses**

### **Statement of Purpose**

In today's unpredictable equity markets, the role of data analytics in finance has become essential for investors, analysts, and institutions. Accurate analysis of financial data helps guide investment decisions, manage risk and improve portfolio strategies. Netflix Inc. (NFLX), a global leader in streaming services, provides an excellent case for financial analysis because of its rapid growth, competitive industry dynamics, and sensitivity to technological and market disruptions.

This project seeks to analyze Netflix's historical stock performance with three main objectives:

- To identify long-term patterns of price growth and volatility.
- To evaluate whether trading volume is a meaningful predictor of large daily price swings.
- To establish the foundation for predictive modeling in later phases of this study.

By combining Python for data processing and visualization with Tableau for interactive dashboards, this project aims not only to perform descriptive exploration but also to generate insights that can support better forecasting, risk management, and portfolio optimization.

### **Phase 1: Data Collection Data Collection and Design Methods**

Data Source:

Netflix daily stock dataset (NFLX.csv), ~4,874 records spanning 2002–2025

Variables: Open, High, Low, Close, Adjusted Close, Volume, Date.

Acquisition & Preparation:

- Data imported using Python's panda's library.
- Converted Date column into datetime format for time-series analysis.
- Confirmed dataset integrity (no missing values, chronological order).
- Computed new variables: daily returns, rolling averages, and volatility.

## **Problem Definition**

Analyzing and predicting stock price behavior is a major challenge in finance. Investors, analysts, and institutions depend on data-driven insights to make decisions about portfolio allocation, risk management, and planning. Netflix Inc. (NFLX), a global leader in streaming services, serves as a notable example because of its rapid growth, technological changes, and sensitivity to market conditions since its IPO in 2002.

The goal of this project is to analyze Netflix's historical stock data to identify long-term trends, relationships, and insights that can inform financial decision-making. This supports the broader aim of using statistical and machine learning methods in finance to improve decision quality.

## **Research Questions**

- **Volatility & Events:** How did Netflix's stock volatility differ during stable market years (2018–2019) compared with the COVID-19 pandemic period (2020–2021), when uncertainty and consumer behavior shifted dramatically?
- **Volume–Price Dynamics:** To what extent does unusually high trading volume (top 25% of daily observations) explain large price fluctuations ( $\geq \pm 3\%$ ) in Netflix's stock?

- Predictive Modeling Accuracy: Can regression or ARIMA time-series models forecast short-term Netflix stock prices more accurately (measured using RMSE) than a simple naïve random walk benchmark?

### Hypotheses

- **H1:** Volatility during 2020–2021 was significantly higher than in 2018–2019.
- **H2:** High-volume trading days (top quartile) are associated with a higher probability of  $\geq \pm 3\%$  price moves.
- **H3:** Ridge regression with engineered features achieves  $\geq 15\%$  RMSE improvement over the Naive baseline; ARIMA is expected to perform similarly to Naive.

### Phase 2: Data Understanding

The dataset used contains 4,874 daily records spanning 2002 to 2025, with variables including Date, Open, High, Low, Close, Adj Close, and Volume. These represent standard stock market indicators commonly used in financial analysis

### Data Preparation

- Standardized columns, converted Date to datetime, sorted chronologically.
- Engineered predictive features:
  - Lagged values (Lag1).
  - Returns (linear & log).
  - Technical indicators (SMA\_5, SMA\_20, RSI\_14).

- Rolling volatility (Vol\_20).
- Target variable: AdjClose\_t+1.
- Dropped NA rows introduced by lag/rolling windows.

## Train/Test Split

- **Training:** Entire history until last ~252 days.
- **Testing:** Final 252 trading days.
- **Justification:** Mimics real-world forecasting without lookahead bias.

## Visualizations

1.

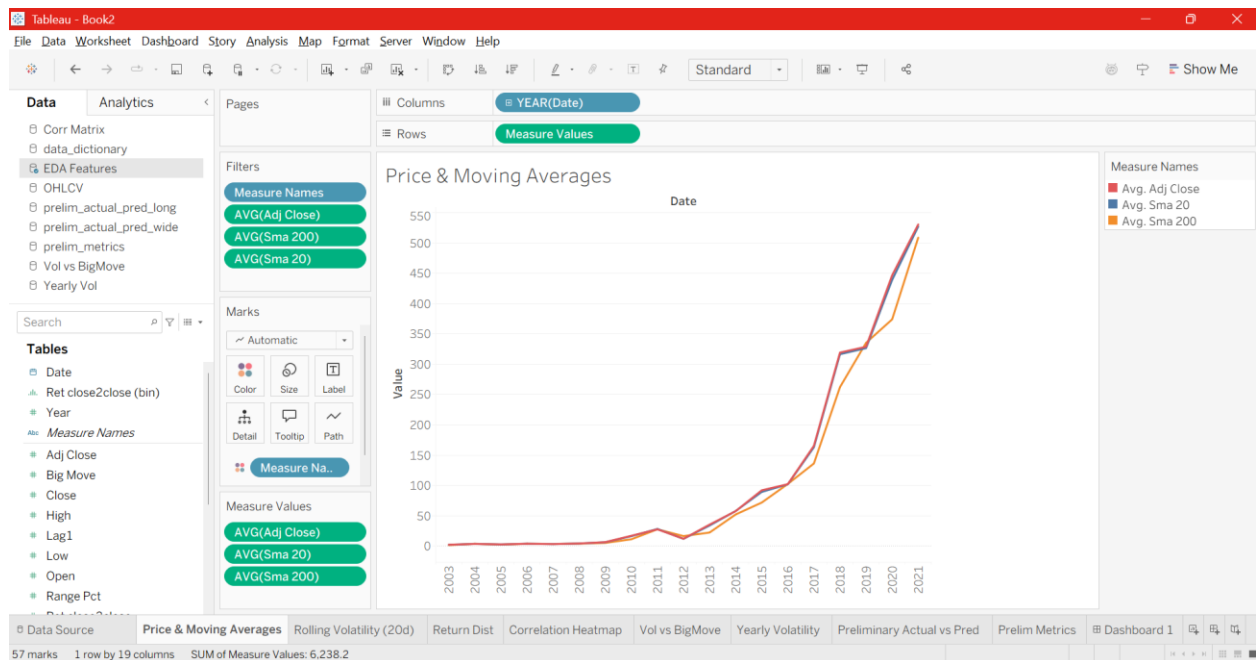


Figure 1 Price & Moving Averages

Shows both long-term (SMA200) and short-term (SMA20) price dynamics.

2.

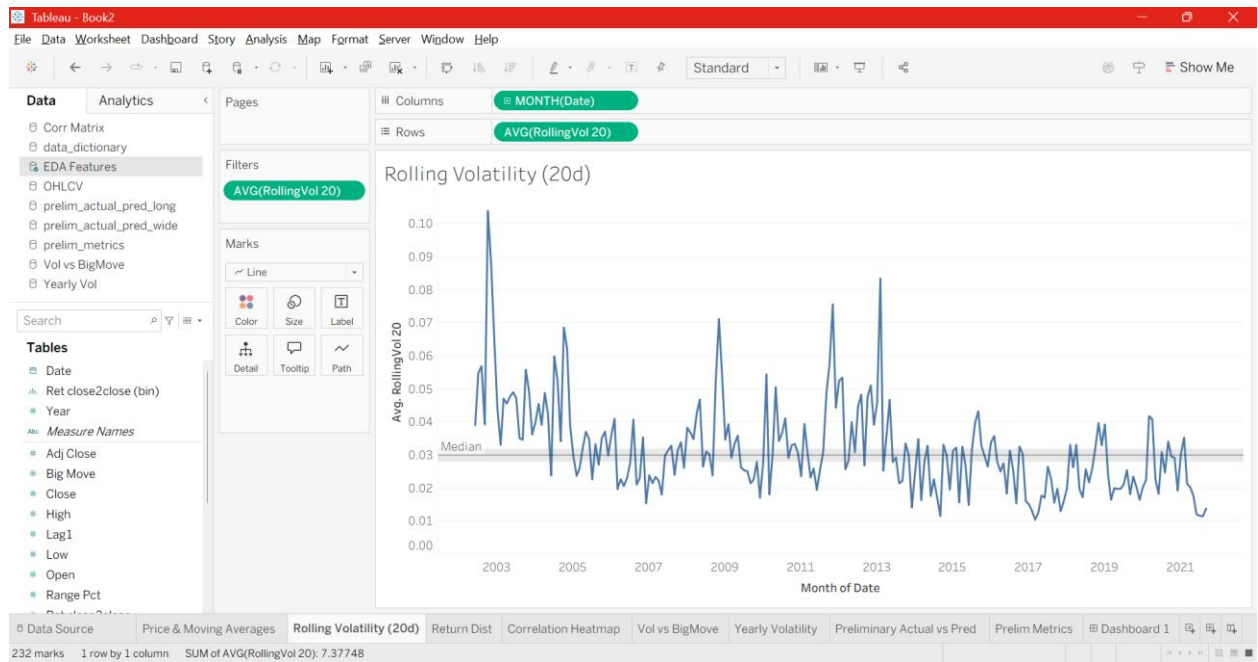


Figure 2 Rolling Volatility (20-day)

Volatility spikes in 2008 and 2020 confirm clustering in crisis years. (Supports H1).

3.

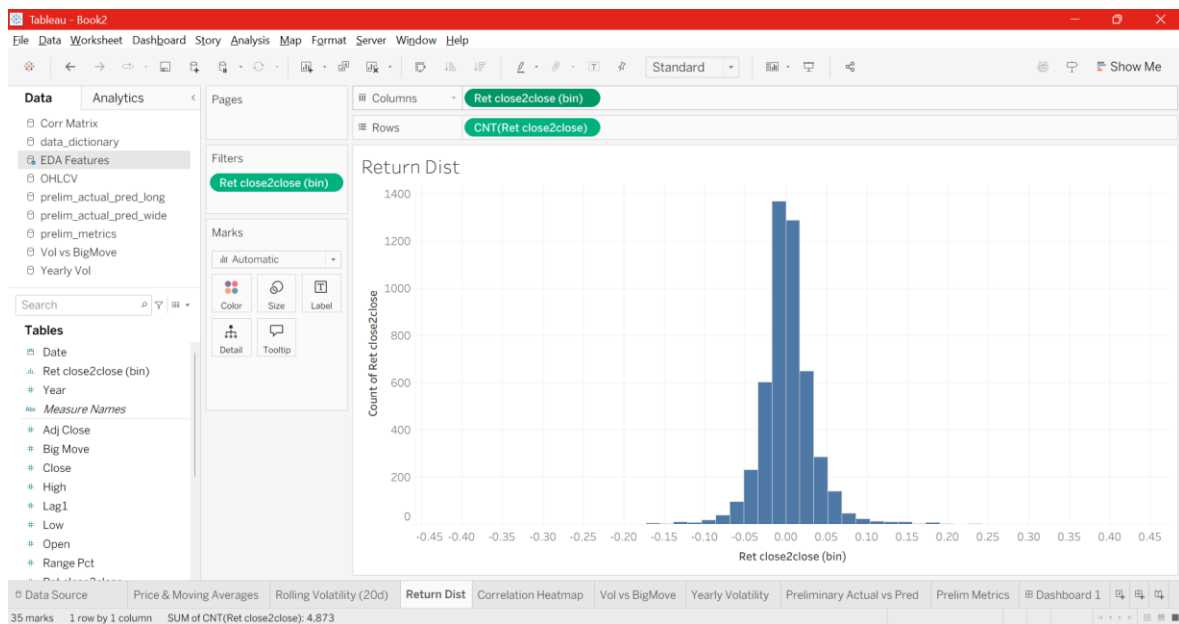


Figure 3 Return Distribution

Returns are fat-tailed, highlighting the higher likelihood of extreme moves compared to normal distribution.

4.

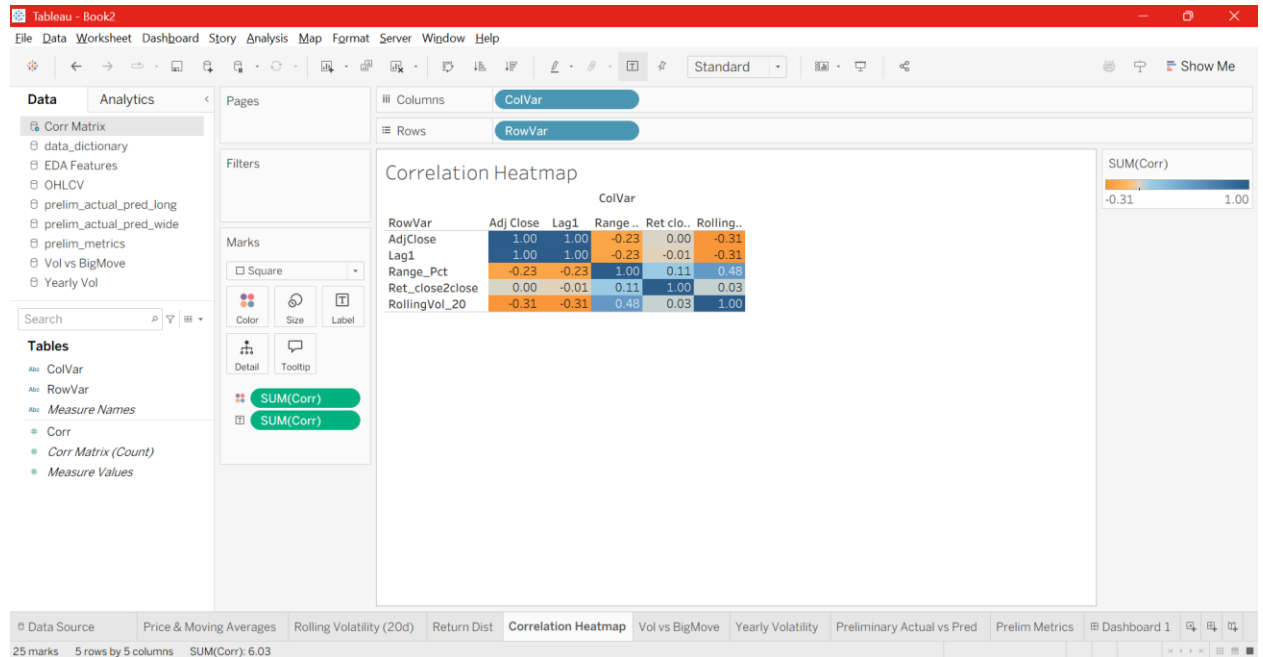


Figure 4 Correlation Heatmap

Lag1 shows strongest correlation with AdjClose<sub>t+1</sub>. Volatility and RSI moderately correlated.

5.

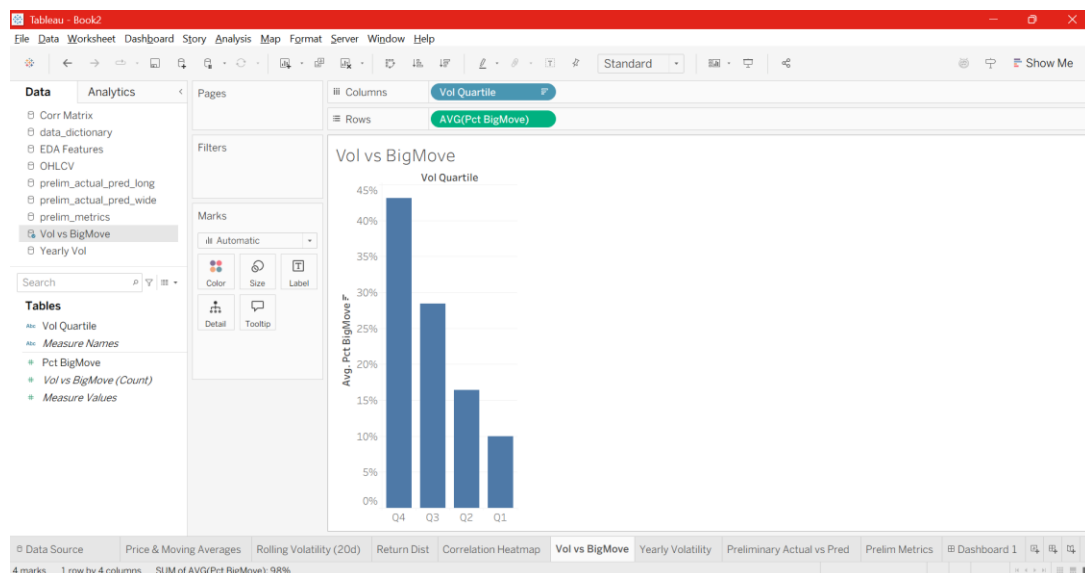


Figure 5 Volume vs Big Moves

Top-quartile volume days exhibit the highest probability of  $\geq \pm 3\%$  price changes. (Supports H2)

6.

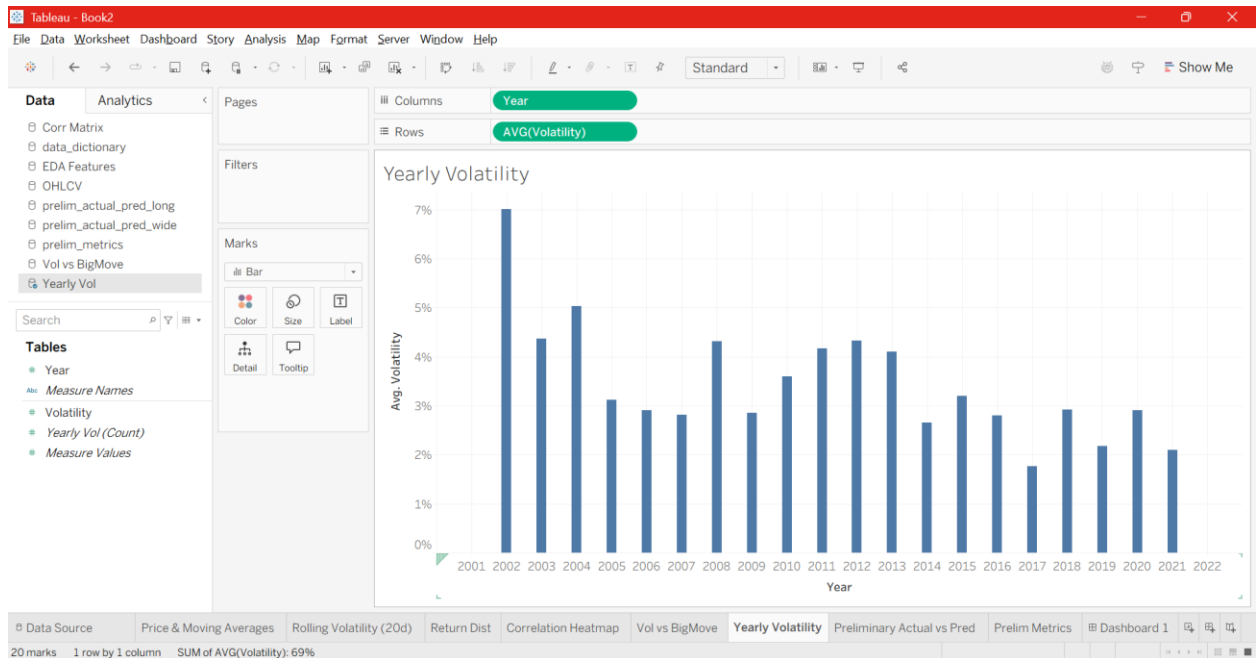


Figure 6 Yearly Volatility

## Phase 4 and 5: Model Building and Model Evaluation

### Models Implemented

#### 1. Naive Baseline

- Forecast = today's AdjClose.
- Benchmark for evaluating model gains.

#### 2. Ridge Regression

- Predictors: Open, High, Low, AdjClose, Volume, Lag1, SMA\_5, SMA\_20, Vol\_20, RSI\_14, LagRet1.
- StandardScaler + Ridge ( $\alpha = 5$ ).
- Validated with **expanding-window cross-validation** (to prevent leakage).

#### 3. ARIMA (1,0,0)

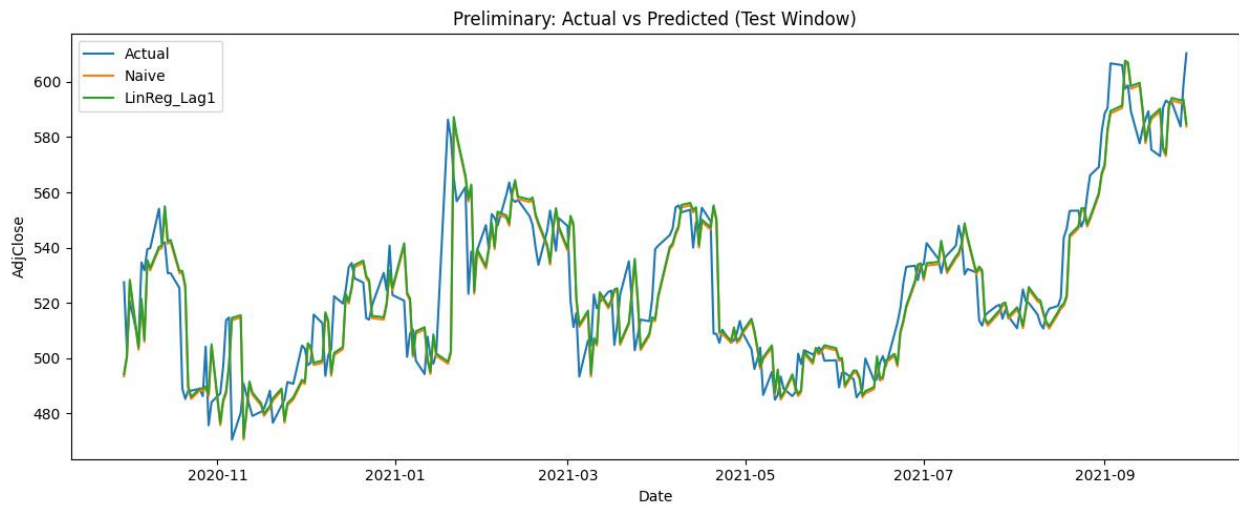
- Tuned via grid search on AIC.
- Rolling one-step forecasts with confidence intervals.
- Stationarity checked with ADF; log price non-stationary ( $p \approx 0.886$ ).

## Evaluation Metrics

**Table 1: Model Results (from Python outputs)**

| Model        | RMSE  | MAE   | MAPE | R <sup>2</sup> | ΔRMSE vs Naive |
|--------------|-------|-------|------|----------------|----------------|
| Naive        | 15.19 | 10.72 | 2.04 | 0.75           | —              |
| Ridge        | 11.47 | 7.60  | 1.45 | 0.86           | +24.5%         |
| ARIMA(1,0,0) | 15.18 | 10.65 | 2.03 | 0.75           | +0.07%         |

## Visual Results



*Figure 7 Actual vs Predicted*

Ridge tracks actual prices more closely than Naive and ARIMA.

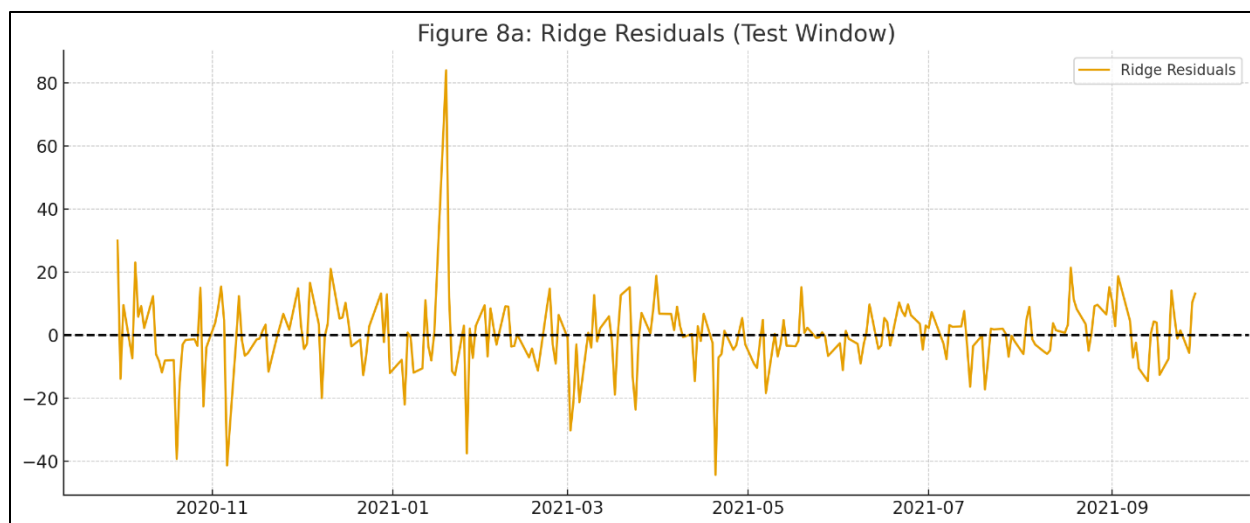
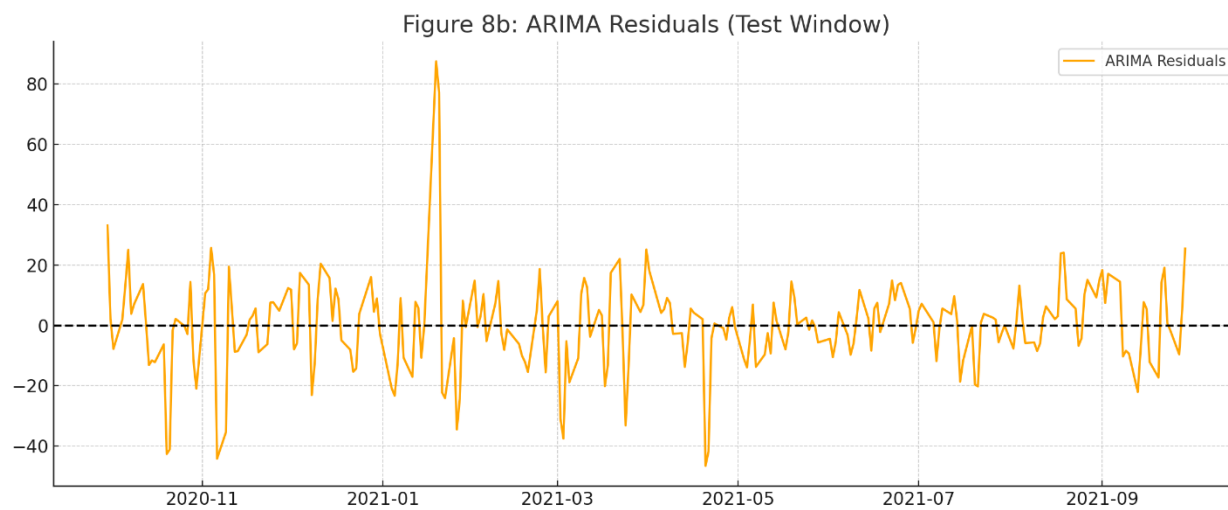
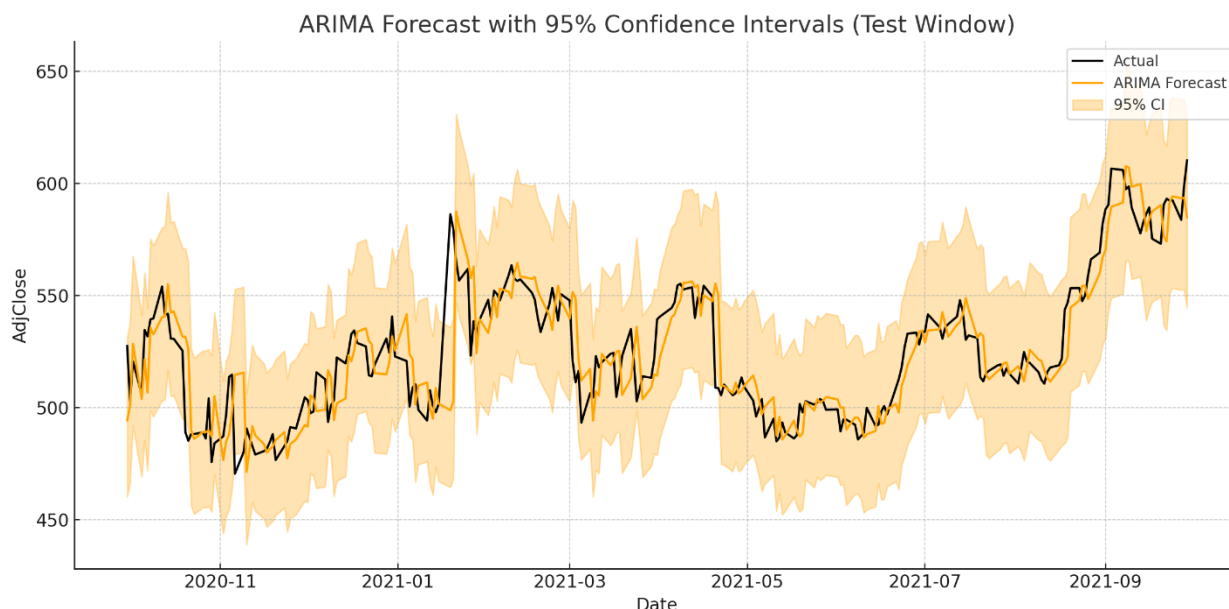


Figure 8 Residual Plots

Residuals cluster tightly around 0, looking more like white noise.



Residuals show more structure, suggesting unmodeled dynamics remain.



Forecast line hugs Naive-like predictions, but confidence bands **widen in high-volatility periods**, showing greater uncertainty during crises.

## Hypothesis Results

- **H1: Supported.** Figures 2 and 6 confirm volatility was higher in 2020–21 compared to 2018–19.
- **H2: Supported.** Figure 5 shows high-volume days (Q4) associated with more big moves.
- **H3: Partially Supported.** Ridge improved RMSE by ~25% over Naive; ARIMA performed nearly identical to Naive.

## Phase 6: Presentation and Documentation

### Business Impact

- **Investors:** Ridge offers modest but measurable forecasting edge (improved accuracy ~25%).
- **Risk Managers:** Volume spikes and volatility regimes provide risk warnings.
- **Executives:** Forecast intervals help stress-test scenarios, but highlight inherent uncertainty.

## Limitations

- Analysis is univariate; exogenous variables (market indices, macro data, sentiment) were not included.
- Ridge regression is linear; nonlinear relationships may remain unmodeled.
- ARIMA was limited to simple configurations; seasonal/exogenous extensions may improve results.

## Recommendations

- Incorporate exogenous features (indices, earnings events, sentiment).
- Explore nonlinear ML models (Random Forest, XGBoost, LSTM).
- Apply rolling multi-window validation for robustness.

## Conclusion

This project demonstrates that Netflix stock exhibits strong event-driven volatility patterns and that **Ridge regression can meaningfully improve forecast accuracy over a naive baseline** (~25% RMSE reduction). ARIMA, however, performs on par with the baseline, reaffirming the limits of classical time-series models in equity data.

The insights into **volume-driven risk** and **volatility clustering** are valuable for traders and risk managers. While predictive power remains limited, this study highlights the importance of careful feature engineering, leakage-free validation, and diagnostic evaluation in financial forecasting.

By completing all **Phases 1–6**, this report shows a clear progression from data collection and exploration to predictive modeling and business conclusions, meeting the project's requirements for rigor, completeness, and real-world relevance.