**NAIROBI STOCK EXCHANGE PREDICTION**

**CRISP-DM**

**AUTHORS:**

Brenda Mutai - Group Lead

Justin Mbugua

Sharon Momanyi

Stephen Munyiala

# **Table of Contents**

[**Table of Contents** 2](#_Toc197303200)

[**CHAPTER 1. INTRODUCTION.** 3](#_Toc197303201)

[**CHAPTER 2. BUSINESS UNDERSTANDING.** 4](#_Toc197303202)

[**CHAPTER 3. DATA UNDERSTANDING.** 6](#_Toc197303203)

[**CHAPTER 4. DATA PREPARATION**. 7](#_Toc197303204)

[**CHAPTER 5. EXPLORATORY DATA ANALYSIS.** 10](#_Toc197303205)

[**CHAPTER 6. FEATURE ENGINEERING** 22](#_Toc197303206)

[**CHAPTER 7. MODELING** 24](#_Toc197303207)

[**CHAPTER 8. MODEL EVALUATION.** 30](#_Toc197303208)

[**CHAPTER 9. DEPLOYMENT** 32](#_Toc197303209)

[**CHAPTER 10. CONCLUSIONS AND RECOMMENDATIONS.** 33](#_Toc197303210)

# **CHAPTER 1. INTRODUCTION.**

This project focuses on analyzing stock price movements for all publicly listed companies on the Nairobi Securities Exchange (NSE) over the years 2013 and 2024. By leveraging daily trading data—including stock prices, trading volume, sector classifications, and other relevant financial indicators—the study aims to uncover patterns and trends that influence market behavior.

Through time series analysis, statistical modeling, and machine learning techniques, the project will evaluate historical price movements to detect seasonal trends, price volatility, and sector-wide correlations. Additionally, outlier detection will be applied to assess extreme market events, ensuring a comprehensive understanding of price fluctuations without disregarding natural market dynamics.

A key focus of the project is to aid investors, financial analysts, policy makers and researchers in identifying potential opportunities and risks within the Kenyan stock market. By generating data-driven insights, the study will help stakeholders make informed decisions regarding portfolio diversification, market entry strategies, and investment planning.

Moreover, forecasting models—including ARIMA, XGBoost, and LSTM—will be deployed to predict future stock price movements, offering a quantitative basis for analyzing possible market directions. The study will also investigate trading volumes to understand liquidity trends, identifying stocks that exhibit high investor activity and those that remain relatively illiquid.

Ultimately, this project serves as a comprehensive exploration of Kenya’s stock market dynamics, aiming to enhance financial knowledge, support investment strategies, and contribute valuable insights to market participants.

# **CHAPTER 2. BUSINESS UNDERSTANDING.**

The Nairobi Securities Exchange stands as a cornerstone of Kenya's financial system, providing essential capital market functions that support economic growth and wealth creation. With its diverse product offerings, robust regulatory framework, and ongoing modernization efforts, the NSE offers significant opportunities for both businesses seeking capital and investors looking for returns. As Kenya's economy continues to develop and integrate with regional and global markets, the NSE is well-positioned to play an increasingly important role in the country's financial landscape and broader economic transformation.

The Nairobi Securities Exchange (NSE) serves as a vital platform for raising capital, enabling businesses and the government to secure funding through shares and bonds for growth and operations. It provides diverse investment opportunities for both local and international investors, fostering wealth creation and portfolio diversification. By mobilizing domestic savings and attracting foreign investment, the NSE contributes to Kenya’s economic development while promoting strong corporate governance. Through supply and demand dynamics, it facilitates price discovery, ensuring securities reflect their true market value. Additionally, it enhances liquidity, allowing investors to easily trade their assets.

**2.1 Problem Statement.**

Investors and financial institutions operating on the Nairobi Securities Exchange (NSE) rely on precise information to make strategic decisions regarding stock trades. By forecasting future stock prices and identifying market trends, they can optimize investment strategies, improve portfolio management, and effectively mitigate risks. Such insights empower traders to seize profitable opportunities, avoid potential losses, and enhance overall financial stability.

Additionally, predictive models enable investors to understand market behavior better, adapt to changing conditions, and maintain a competitive edge. With tools for analyzing historical trading data, stakeholders can uncover patterns, assess the impact of external factors like economic policies or global market shifts, and make data-driven choices that align with their financial goals. These advancements are essential for thriving in a dynamic stock market environment like the NSE.

**2.2 Objectives**

The overall aim of this study is to develop time series models to forecast future stock prices.

Specifically, the study seeks to:

1. Provide insights into which stocks might perform well based on historical trends and predictive models, which will allow for more informed decision-making.

2. Offer short-term predictions of stock prices or trends to support timely buy/sell decisions, potentially improving their profitability.

3. Develop a machine learning-based tool that provides predictive insights and visualizations for NSE market trends.

**2.3 Stakeholders**

**The study focuses and addresses the following stakeholders:**

1. Individual Investors: These are everyday people who buy and sell stocks, bonds, or other securities to grow their personal wealth and achieve financial goals.
2. Institutional Investors: Large organizations, such as pension funds, insurance companies, and mutual funds, that invest substantial amounts of money in financial markets.
3. Stockbrokers and Analysts: Stockbrokers facilitate the buying and selling of securities on behalf of investors, while analysts study market trends and company performance to provide investment insights.
4. NSE and Regulatory Bodies: The Nairobi Securities Exchange (NSE) operates the stock market, while regulatory bodies, such as the Capital Markets Authority (CMA), oversee trading activities to ensure fairness and transparency.
5. Financial Advisors: Professionals who offer guidance on investment decisions, helping individuals and businesses optimize their financial strategies based on goals and risk tolerance.

**2.4 Data Understanding.**

The data is a compilation of historical daily stock market price data relates to the Kenyan Nairobi Securities Exchange (NSE) for the years 2013 to 2024. It was sourced from *https://data.mendeley.com/(Kenya Nairobi Securities Exchange (NSE) All Stocks Prices 2023-2024)*

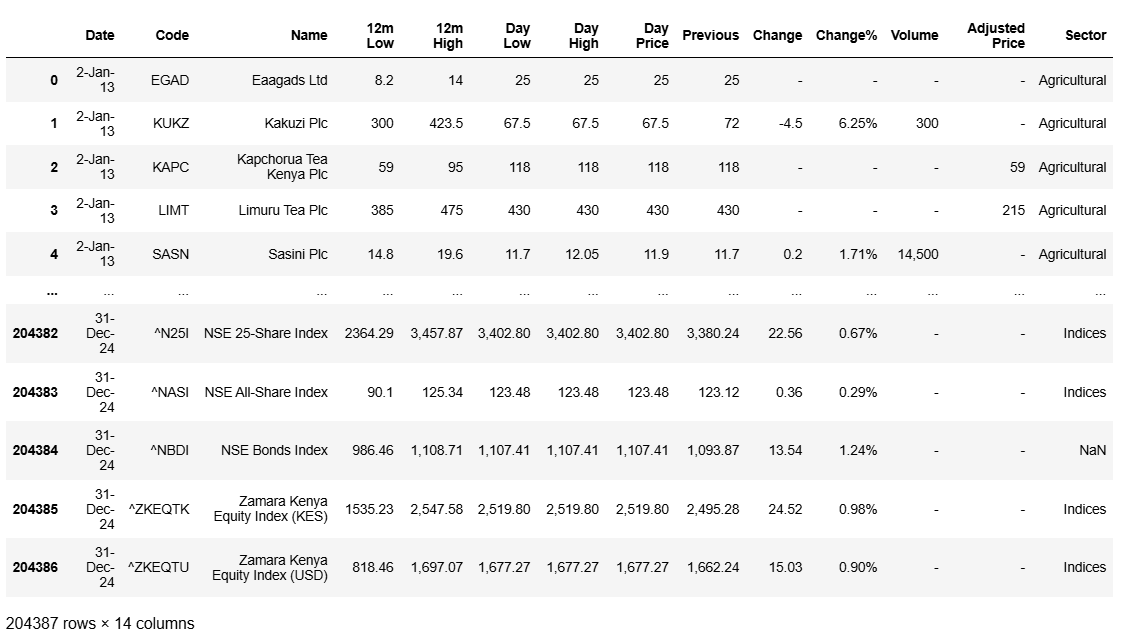
The data was scrapped from a publicly accessible website -<http://live.mystocks.co.ke/> - licensed by NSE by exporting raw web data to spreadsheets, then cleaned up to a final CSV.

Each stock data row has 13 data columns (1) Date (2) Stock Code (3) Stock Name (4) 12-month Low price (5)12-month High price (6) Day's Low price (7) Day's High price (8) Day's Final Price (9) Previous traded price (10) Change in price value (11) Change in price % (12) Volume traded (13) Adjusted price. One additional CSV file is also provided to show the stocks market sector, with three (3) columns as: (1) Market sector (2) Stock Code (3) Stock Name.

**2.5 Metric of success.**

1. Model Performance: Accuracy of stock price forecasts measured by metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) or Mean Squared Error(MSE).
2. Insight Utility: The relevance and actionability of the identified market trends, sector performance analyses, and trading patterns for stakeholder decision-making (Qualitative).
3. Completeness: Successful analysis covering the specified timeframe (2013-2024) and scope (all listed companies, key data points).

# **CHAPTER 3. DATA UNDERSTANDING.**

****

The dataset is structured in a tabular format with 14 columns and 204,387 rows.

Missing Data: "Previous", "Adjusted Price" and “Volume” fields are often empty.

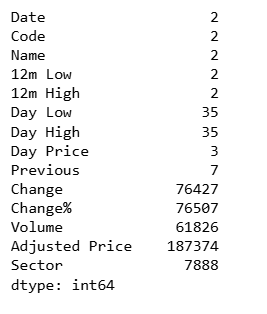
It contains variable such as:

* + **Date**: Trading date (e.g., "03-Jan-2023").
  + **Code**: Stock ticker symbol (e.g., "EGAD", "KUKZ").
  + **Name**: Company or index name (e.g., "Eaagads Ltd", "Kakuzi Plc").
  + **12m Low/High**: 12-month lowest and highest prices.
  + **Day Low/High**: Daily trading range.
  + **Day Price**: Closing price for the day.
  + **Previous**: Previous day's closing price (missing for some entries).
  + **Change/Change%**: Absolute and percentage change from the previous day.
  + **Volume**: Trading volume (some entries missing or zero).
  + **Adjusted Price**: Not populated in the sample.
  + **Sector**: Classification (e.g., "Agricultural").

# **CHAPTER 4. DATA PREPARATION**.

The data appears to have several missing values. Next we checked the dataframe information to find out how many missing values are in each column. This is important when deciding how we will deal with the missing values.

On initial inspection, we observed the data to be clean but after removing the “-” it generated the following features.



The dataset contains missing values across several columns. The 'Previous' column has 7 missing entries. 'Change' and 'Change%' have a significant number of missing values, with 76427 and 76507 respectively. 'Volume' also has a substantial number of missing entries (62826). The 'Adjusted Price' column exhibits the highest number of missing values at 187374. Finally, the 'Sector' column has 7888 missing values. All other columns ('Date', 'Code', 'Name', '12m Low', '12m High', 'Day Low', 'Day High', 'Day Price') have 2,2,2,2,35,35,3 missing values. This highlights there is significant work to do to clean the data.

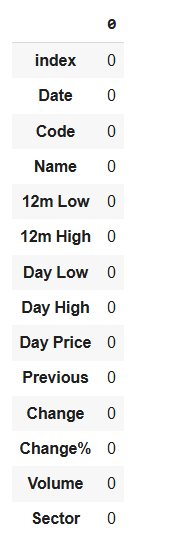
To prepare the numeric columns for analysis, we first need to clean them by removing any commas and spaces before converting them to numeric.

We also have to change the date from string format to datetime. Since the dates in the dataset has the day first while other dates in the dataset has the month first, we used a function that accounts for this.

We then proceeded to do the following to the data.

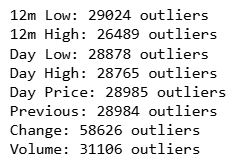
* Drop the **Adjusted Price** column because it doesn't contain any values.
* Replace the missing values in the **Sector** column with 'Unknown'.
* Drop missing values from the **Previous** column which has a very small number of missing values.
* Fill the missing values in the **Change** and **Change%** columns by calculating the values.
* Replace the missing values in the **Volume** column with the median for each stock.

After cleaning this was the final generated clean data.



After data cleaning, 185,336 rows remain, indicating a substantial dataset that has undergone refinement to remove inconsistencies, errors, duplicates, and irrelevant entries. This ensures improved data quality, enhancing the reliability of insights and predictions drawn from the cleaned dataset.

We analyzed the dataset for outliers to gain deeper insights, but given the inherent volatility of financial data, we have chosen to retain them. Most outliers reflect natural market fluctuations rather than errors, making them valuable for understanding real-world trends.



Outlier Report: Stock prices, particularly intraday highs, lows, and trading volume, tend to be highly volatile. Sharp price movements (reflected as "Change") and sudden surges in volume frequently occur during major market events, news releases, or heightened trading activity. While these may appear as statistical outliers, they often carry valuable signals for predictive modeling.

The fact that the "Change" feature has the most "outliers" (58,626) is especially significant. Price fluctuations typically exhibit heavy-tailed behavior, meaning extreme values occur more frequently than they would in a normal distribution. These large price movements, labeled as outliers, are often precisely what prediction models aim to capture, not eliminate.

# **CHAPTER 5. EXPLORATORY DATA ANALYSIS.**

We checked the data description to identify the numerical columns.



**Day Price:** The average 'Day Price' is approximately 102.74, with a median (50th percentile) of 17.15. The prices range significantly, from a minimum of 0.16 to a maximum of 3330. The standard deviation of 283.99 suggests a wide distribution of daily prices.

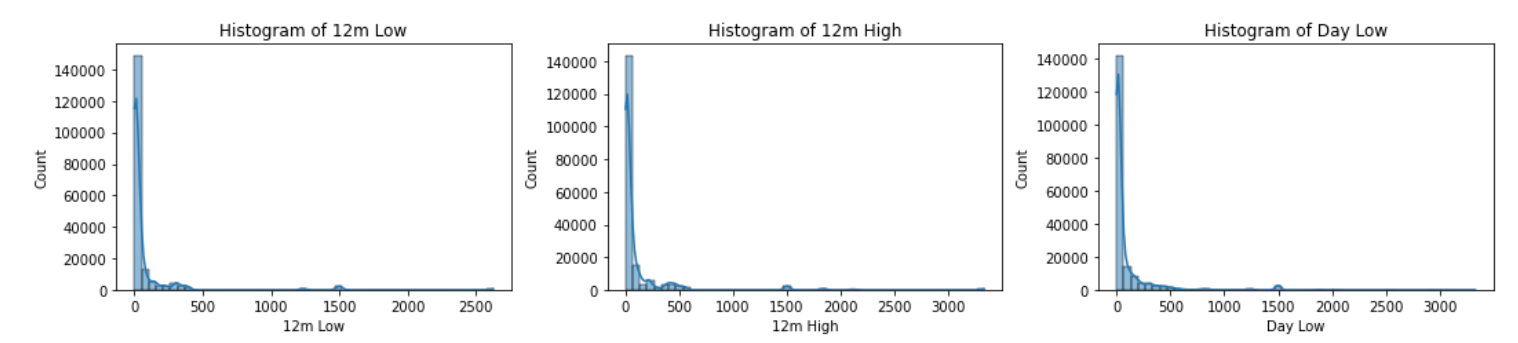
**Change:** The average 'Change' is a small positive value (0.01448), with a median of 0. The changes range from -250.00 to 1438.25, indicating both significant drops and increases in value. The standard deviation is 4.97.

**Change%:** The average 'Change%' is approximately 1.47%, with a median of 0.49%. The range is substantial, from 0% to 2329.15%, and the standard deviation is 6.08, suggesting high volatility in percentage changes.

**Volume:** The average 'Volume' is approximately 4.296 x 10^5, while the median is much lower at 6500. This large difference between the mean and median, along with a maximum volume of 2.085 x 10^9, indicates a highly skewed distribution with a few instances of extremely high volume. The standard deviation is also very large (2.648 x 10^6).

**12m Low and High:** The 12-month low prices range from 0.16 to 2625, with an average of around 77.7. The 12-month high prices range from 0.23 to 3330, with an average of around 104.76.

**5.1 Univariate Analysis**



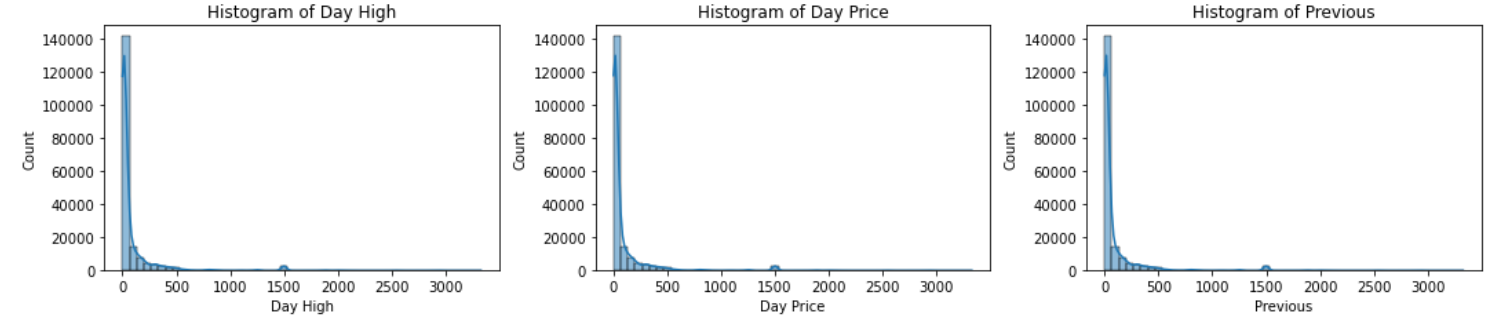
The histograms for '12m Low', '12m High', and 'Day Low' exhibit a strong positive skew.

**Majority of values are concentrated at the lower end of the price range.** A large number of observations have relatively low 12-month lows, 12-month highs, and daily low prices.

**A long tail extends towards higher values.** There are fewer observations with significantly higher prices, creating the right-skewed distribution.

**Potential for outliers on the higher end.** The extended tail suggests the presence of some extreme high values in these price features.

In essence, these price variables are not normally distributed, with a bias towards lower values and occasional spikes into higher price ranges.

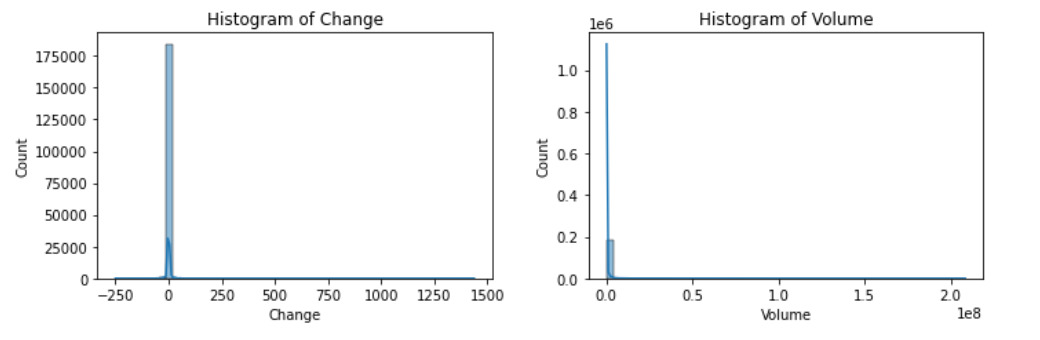


**Lower price ranges are more frequent:** The majority of the data points for the day's high, the closing price, and the previous day's closing price are clustered at the lower end of the price spectrum.

**Higher prices are less common:** The distributions have long tails extending towards higher values, signifying a smaller number of trading instances with significantly higher prices.

**Potential for right-skewed modeling:** Statistical models for these features might need to account for this non-normal distribution.

The consistent skew across these related price metrics suggests a common characteristic of the trading data where lower-priced assets or lower price movements are more prevalent.

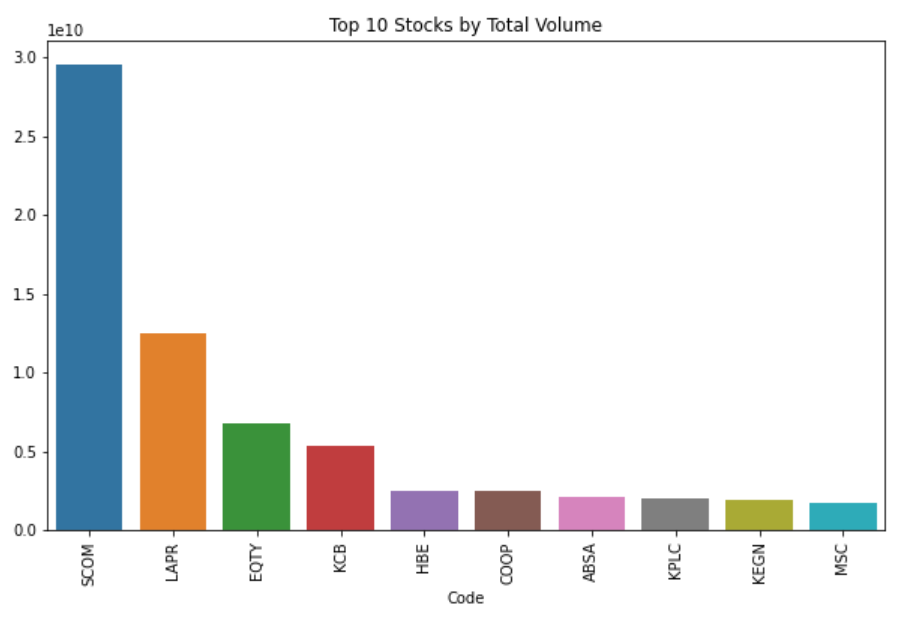


**Histogram of Change:** This distribution is heavily concentrated around zero, indicating that the most frequent price change is no change. There are smaller frequencies of both positive and negative changes, with the range extending from approximately -250 to +300. The distribution appears somewhat symmetrical around zero, although the central spike dominates.

**Histogram of Volume:** This distribution exhibits a strong positive skew. The vast majority of trading volumes are low, clustered near zero. There is a long tail extending to the right, indicating that while infrequent, there are instances of extremely high trading volumes, reaching up to 2 x 10^8. This suggests that most trading days involve relatively low volume, with occasional periods of very high activity.

**5.2 Bivariate Analysis**

1. Top 10 Stocks by Volume and Price Change

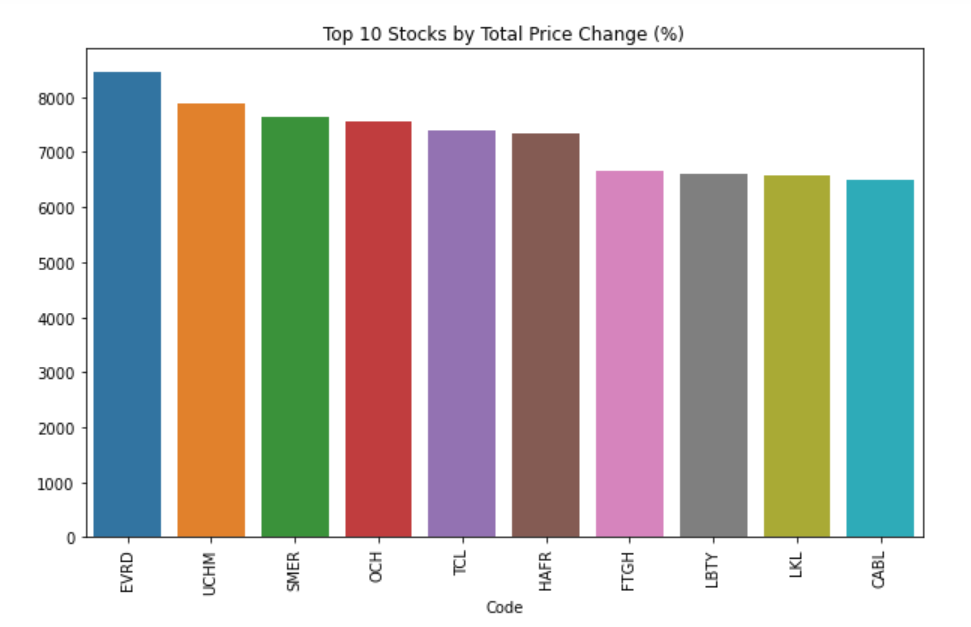


**Top 10 Stocks by Total Volume:**

The most striking observation is the overwhelming dominance of the stock SCOM. Its total volume is approximately 2.9 x 10¹⁰ (29 billion), which is more than double the volume of the second-ranked stock, LAPR (approximately 1.2 x 10¹⁰ or 12 billion). This indicates a highly concentrated trading activity, with SCOM being an exceptionally liquid or actively traded stock compared to all others in the top 10.

Significant Drop-off: There's a steep decline in total volume after "SCOM" and "LAPR."

**Top 10 Stocks by Total Price Change (%):**



All stocks in the top 10 list show incredibly high positive price changes, indicating significant appreciation over the period analyzed. The lowest change among these top performers is approximately 6,500%, while the highest approaches 8,300%.

"EVRD" is the top performer, registering a total price change of roughly 8,300%.

There is a relatively close clustering of performance among the top few stocks. "UCHM," "SMER," "OCH," "TCL," and "HAFR" all exhibit price changes in the range of 7,300% to 7,900%, showing strong and comparable growth.

1. Total volume by sector

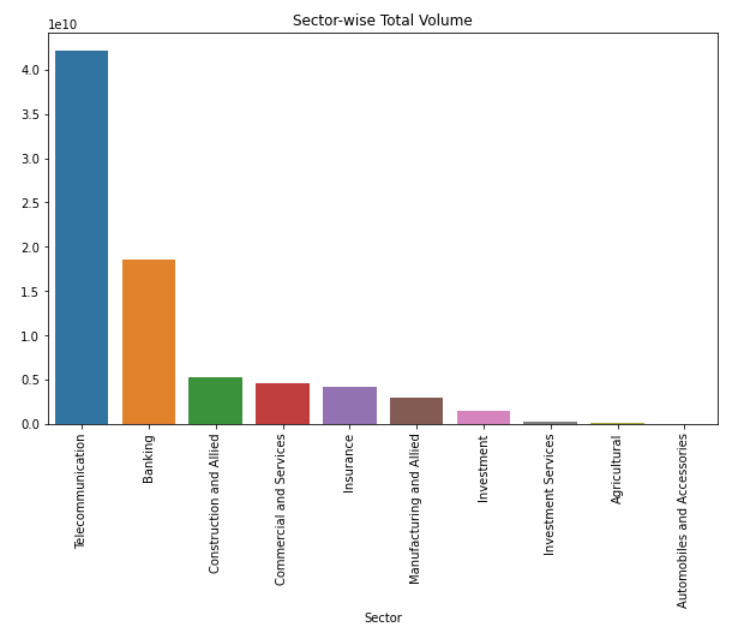
Telecommunication Dominance: The Telecommunication sector exhibits an overwhelmingly high total trading volume, reaching approximately 42 billion units. This volume is more than double that of the next highest sector.

Banking Sector's Significance: The Banking sector secures the second position with a substantial total volume of about 18.5 billion units, highlighting its considerable market activity.

The "Investment," "Investment Services," "Agricultural," and "Automobiles and Accessories" sectors show significantly lower trading volumes, ranging from about 1.5 billion down to just 0.2 billion, indicating relatively less market activity in these areas.

Market liquidity and investor interest are highly concentrated within a few key sectors.

The Telecommunication sector dominates in terms of overall trading activity, with a volume significantly higher than any other sector. The Banking and Commercial & Services sectors also show substantial trading interest. The remaining sectors have considerably lower total trading volumes in comparison.



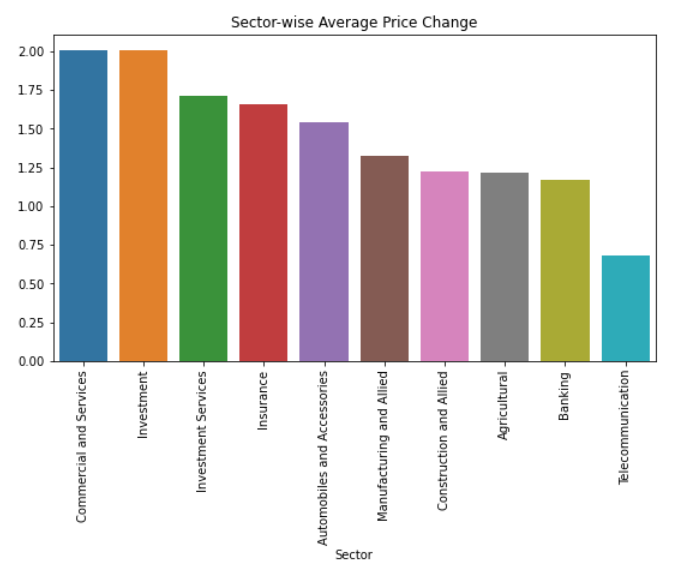
1. Sector wise average price change

The "Commercial and Services" and "Investment" sectors demonstrate the highest average price change, both registering approximately 2.0 units. This indicates strong positive performance on average within these sectors.

Sectors such as "Investment Services," "Insurance," "Automobiles and Accessories," and "Manufacturing and Allied" show moderate average price changes, ranging from roughly 1.7 down to 1.3.

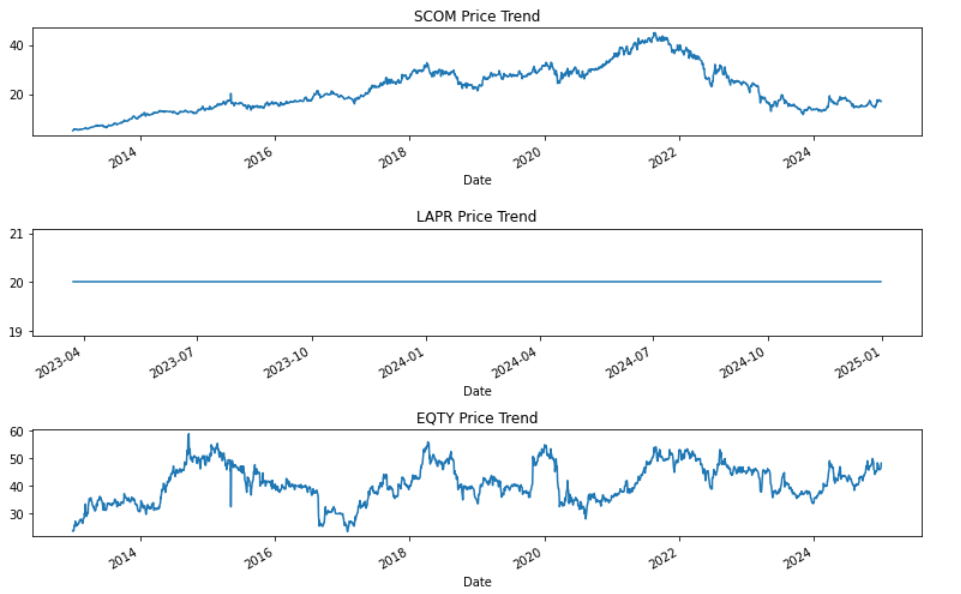
Conversely, the "Telecommunication" sector shows the weakest average price change. This differentiation highlights the diverse dynamics and investment opportunities or risks present across different market sectors.

The Investment sector shows the most significant average upward price movement. Interestingly, the Telecommunication sector, which had the highest total trading volume, displays the lowest average percentage price change among all listed sectors. This suggests that high trading volume doesn't necessarily translate to large average price increases.



1. Price and Volume Trends Over Time

Price trend

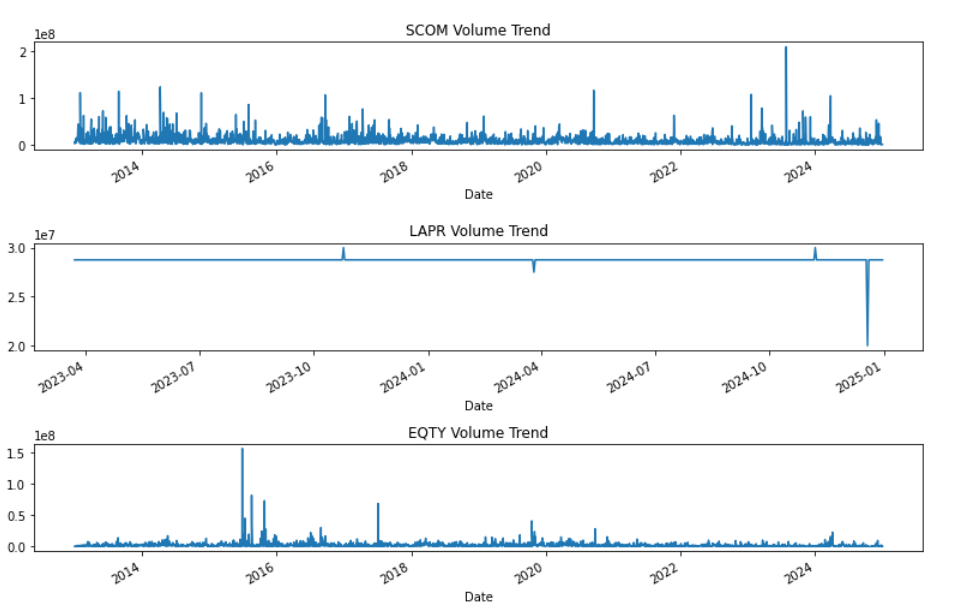


Long-term Growth & Correction: From early 2013, SCOM experienced a steady and significant upward price trend, peaking around mid-2022 at approximately 42 units. The stock has shown considerable volatility, particularly during its growth phase and subsequent decline.

Over the period shown (early 2023 to early 2025), LAPR's price trend is remarkably flat, consistently holding at approximately 20 units. This extreme stability suggests either very low trading activity, a fixed pricing mechanism

EQTY's price trend over the period from early 2013 to early 2025 is characterized by high volatility and cyclical patterns. The price fluctuates significantly, generally within a range of 30 to 60 units.

1. Volume trend

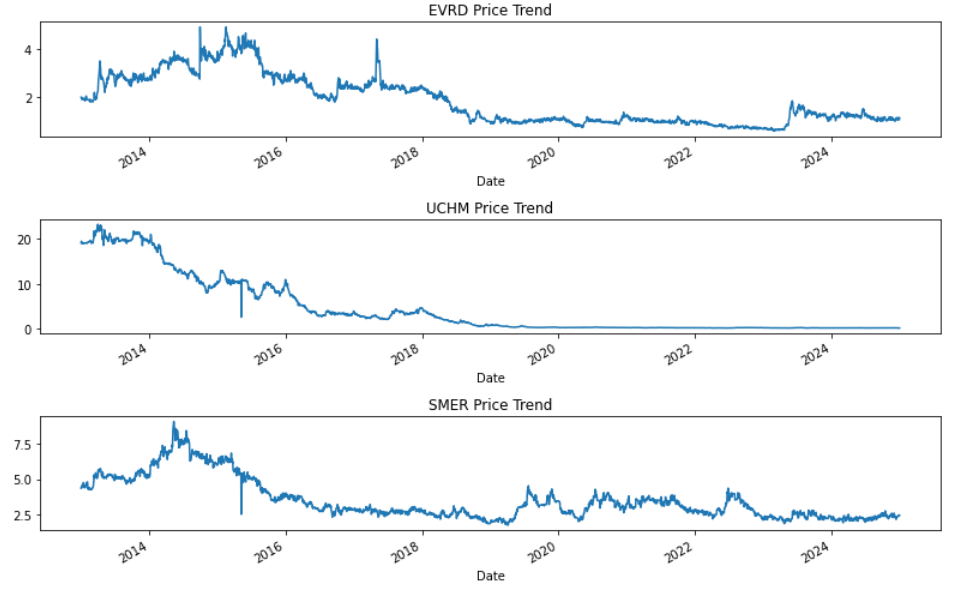


Consistent Activity with Spikes:SCOM over the period from early 2013 to early 2025, SCOM generally maintains a baseline trading volume, but it is frequently punctuated by notable and sometimes very significant spikes in volume. These spikes indicate bursts of intense trading activity.

From early 2023 to early 2025, LAPR exhibits an unusually flat and stable trading volume, consistently hovering around 2.8 x 10⁷ units for most of the period.

EQTY's volume trend from early 2013 to early 2025 shows a generally lower average trading volume compared to SCOM, but it also experiences significant, albeit less frequent, spikes.

1. Price change trend

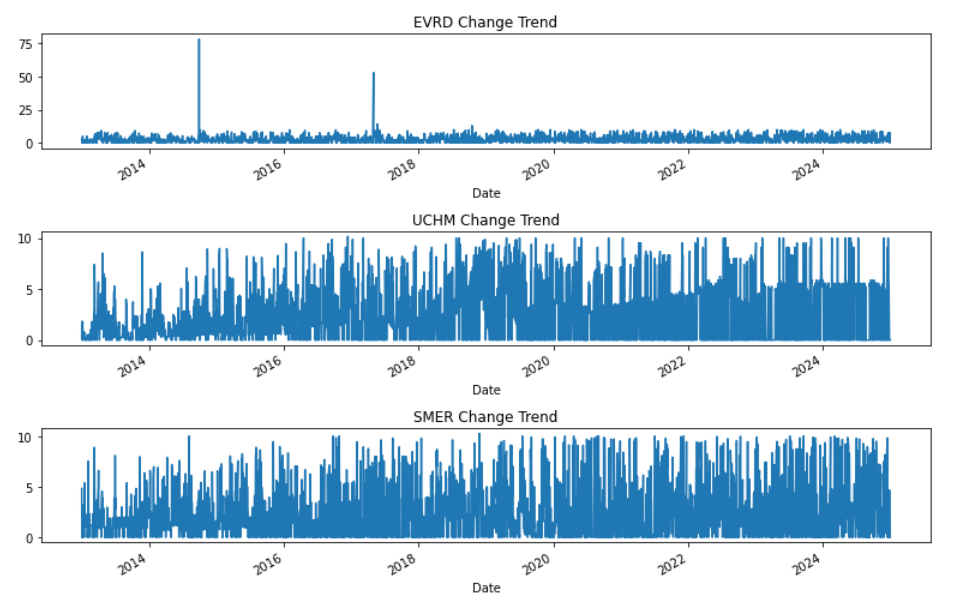


EVRD experienced a strong upward trend from 2013, peaking around mid-2015 at approximately 4.5 units. Subsequently, it embarked on a prolonged downward trend, with intermittent recoveries, settling at a much lower price level, fluctuating between 1.0 and 2.5 units for most of the later period.

UCHM shows a stark and consistent downtrend following an initial peak of over 20 units in early 2014. The price experienced a sharp decline, falling below 1.0 unit by late 2018.

SMER reached a peak of approximately 8.0 units in late 2014, followed by a substantial decline to around 2.5 units by mid-2016.

1. Volume change trend

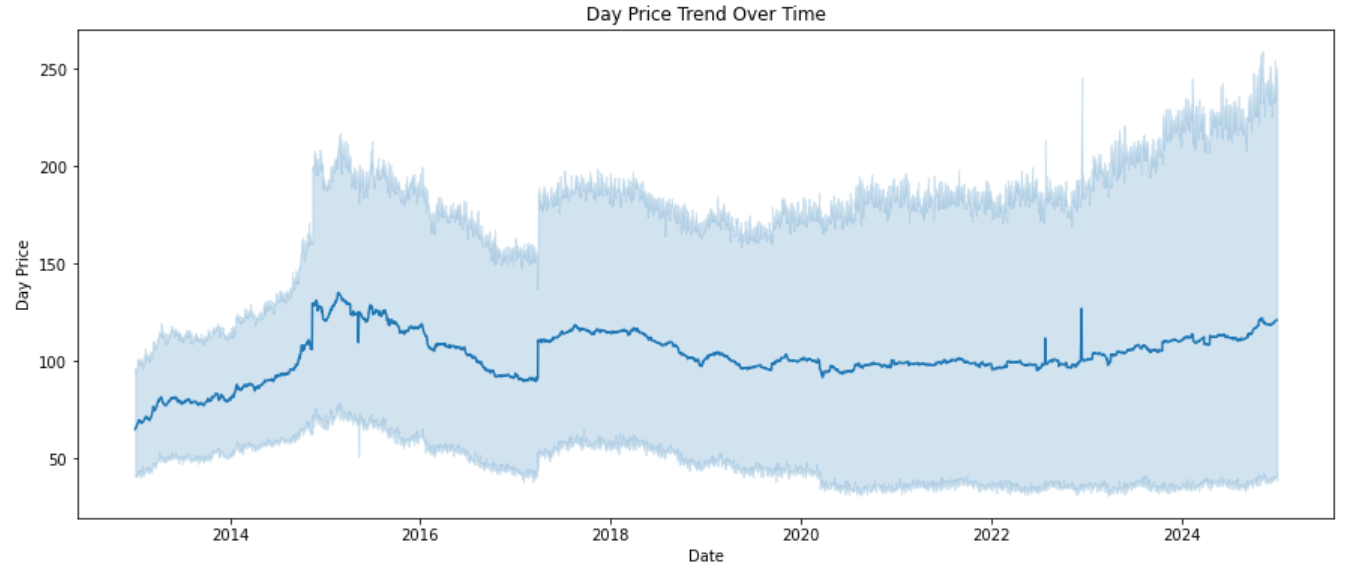


EVRD's daily price changes are typically small, mostly fluctuating between 0 and 10 units. However, the stock is prone to rare but significant positive spikes, with "Change" values reaching up to approximately 75 units. These sharp spikes indicate isolated periods of very large price movements.

UCHM experiences considerable price movement on a day-to-day basis. Given its previously observed long-term price decline, this implies significant daily absolute changes, regardless of direction.

SMER also undergoes significant price movements on a regular daily basis, indicative of a highly active and volatile trading environment.

1. Day Price trend over Time.

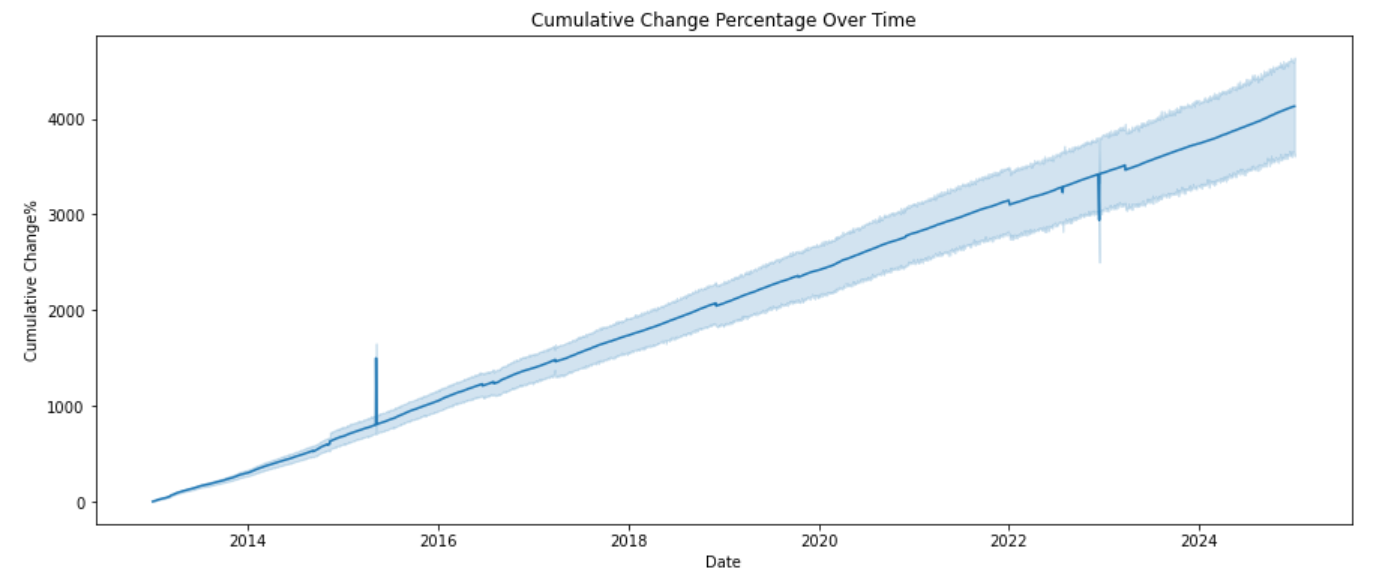


Initial Growth and Subsequent Decline (2013-2017): The "Day Price" showed a strong upward trend from early 2013, peaking around mid-2015 with prices reaching approximately 130 units on average. Following this peak, there was a noticeable decline until mid-2017, where the average price settled around 90 units.

Period of Relative Stability with Fluctuations (2017-2023): From mid-2017 to mid-2023, the average "Day Price" entered a phase of relative stability, generally fluctuating within a band of approximately 90 to 120 units. While stable, there were noticeable short-term price spikes within this period, indicating occasional bursts of higher average prices.

Recent Uptick (2023-2025): From mid-2023 onwards to early 2025, there appears to be a moderate upward trend, with the average price rising from around 100 to 120 units.

1. Cumulative Change Percentage.



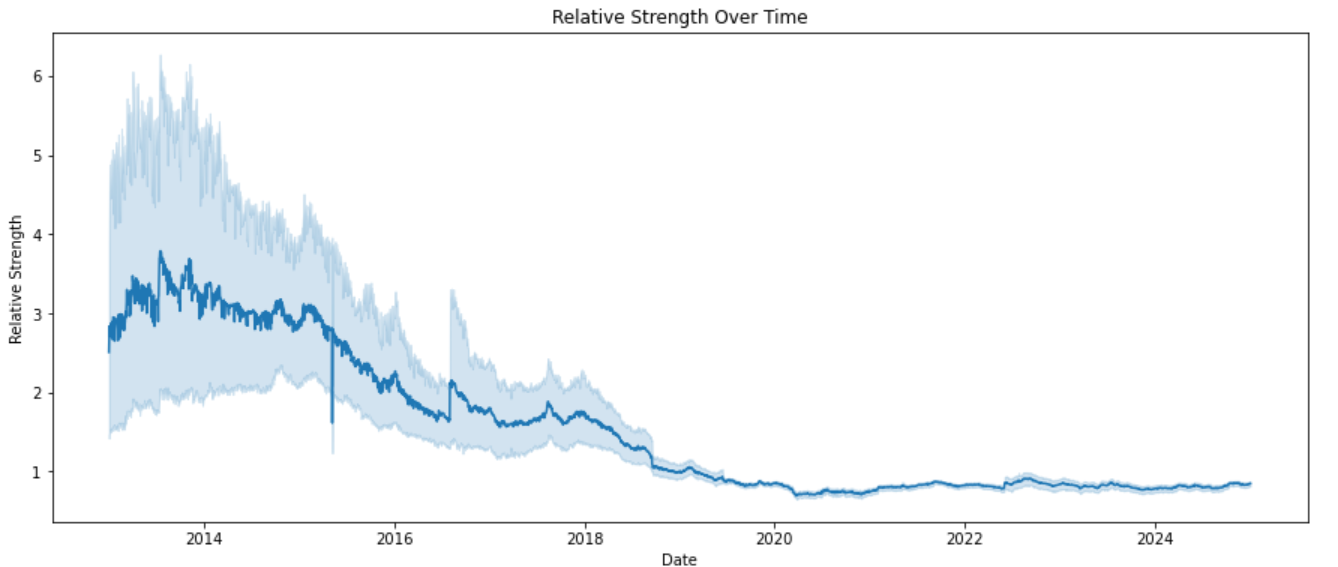
Consistent Upward Trend: The cumulative change percentage shows a remarkably consistent and strong upward trend throughout the entire period. Starting near 0% in early 2013, it steadily climbs to reach over 4000% by early 2025. This indicates a sustained period of positive cumulative growth.

Steady Rate of Change: The slope of the main trend line appears relatively constant, suggesting a fairly steady rate of positive cumulative change over the long term, rather than periods of accelerating or decelerating growth.

While the overall trend is smooth, two distinct periods show a temporary sharp increase in the main trend line and a wider spread in the shaded area: one around late 2015 / early 2016 and another around late 2022 / early 2023. These could represent periods of accelerated growth or increased volatility.

A robust and sustained long-term positive cumulative change, indicating consistent overall growth over the past decade.

1. Relative Strength over Time.



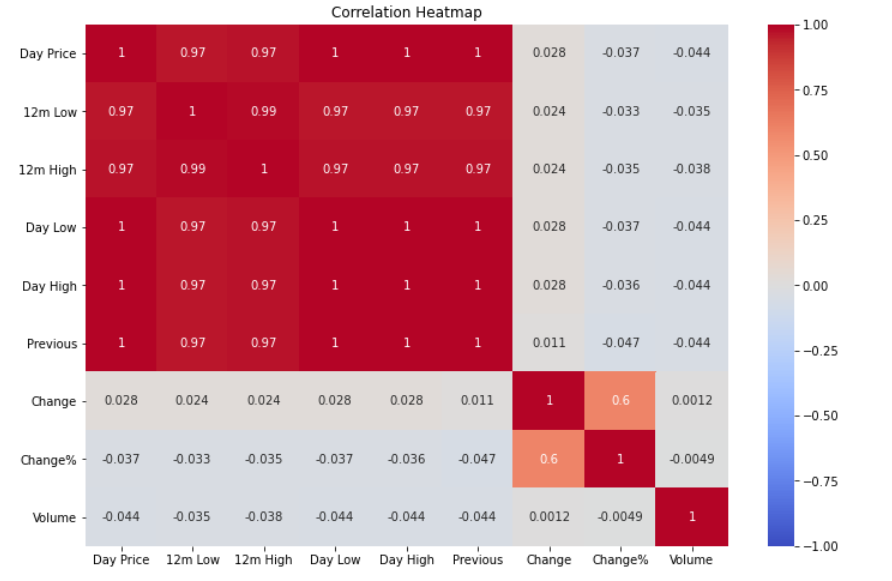
Early Strong Performance (2013-2014): In the initial period, from early 2013 to mid-2014, the Relative Strength was high, consistently fluctuating between approximately 3.0 and 3.5. This indicates that the "Day Price" was frequently trading significantly above its previous 12-month highs, suggesting strong positive momentum or a sustained upward price trend in which new 12-month highs were being consistently surpassed. This period also showed high volatility in this metric.

Sustained Decline (2014-2020): Following this strong period, there was a clear and consistent decline in Relative Strength from mid-2014, eventually dropping below 1.0 by late 2019/early 2020. This signifies that the "Day Price" gradually moved further away from, and eventually consistently below, its 12-month highs, indicating a reversal of momentum and a weakening price trend relative to its recent peaks. The volatility in Relative Strength also decreased during this decline.

Recent Underperformance and Stability (2020-2025): From early 2020 onwards to early 2025, the Relative Strength has remained at a very low and stable level, fluctuating narrowly around 0.8 to 0.9. This indicates that the "Day Price" has consistently traded significantly below its 12-month highs, suggesting a prolonged period of underperformance relative to its annual peaks, with minimal short-term fluctuations in this relative measure.

A notable shift from a period where the asset was frequently setting or exceeding its 12-month highs, to a prolonged period where it consistently trades at levels significantly below those highs, reflecting a sustained decline in its relative strength.

**5.3 Multivariate analysis**



**Extremely Strong Positive Correlation**: Features directly related to stock price levels, including 'Day Price', '12m Low', '12m High', 'Day Low', 'Day High', and 'Previous' price, show exceptionally high positive correlations with each other

**Moderate Positive Correlation**: 'Change' (absolute daily price change) and 'Change%' (daily percentage price change) exhibit a moderate positive correlation of 0.6. This is logical, as percentage change is derived from the absolute change but scaled by the prior day's price, preventing a perfect 1.0 correlation.

**Weak or Negligible Correlations with Price Movements and Volume**:

Price Levels vs. Changes: All the price-related features ('Day Price', '12m Low', '12m High', 'Day Low', 'Day High', 'Previous') show very weak or negligible linear correlations (coefficients close to zero, e.g., 0.028, -0.037) with 'Change' and 'Change%'. This suggests that the absolute price level of a stock has little to no linear relationship with the magnitude or direction of its daily price movements. Volume vs. Other Features: 'Volume' demonstrates extremely weak linear correlations with all other features, including price levels, price changes, and percentage changes (coefficients ranging from -0.044 to 0.0012). This indicates that trading volume is largely linearly independent of stock prices and their daily movements.

# **CHAPTER 6. FEATURE ENGINEERING**

* 1. Moving Averages (SMA and EMA)

We calculate 10-day and 50-day Simple Moving Averages(SMA), as well as the 10-day and 50-day Exponential Moving Averages(EMA), to highlight trends.

**SMA\_10** and **SMA\_50**: Simple moving averages over 10 and 50 trading days.  
**EMA\_10** and **EMA\_50**: Exponential moving averages over 10 and 50 days (more weight to recent prices).

These indicators help smooth out price fluctuations and identify overall market trends, making them essential for detecting potential reversals or continuations in stock movement. While SMA assigns equal weight to all past prices, EMA prioritizes recent price data, making it more reactive to short-term shifts in market conditions.

Moving averages are used for **trend identification, entry/exit points**, and **confirmation of breakout patterns**, helping traders refine their decisions.

They provide a comprehensive view of price trends, from immediate shifts to sustained long-term directions, making them indispensable tools for financial analysis and predictive modeling.

* 1. Relative Strength Index (RSI).

The RSI measures the speed and change of price movements. Values above 70 indicate that a stock is overbought while values below 30 indicate that it's oversold.

The Relative Strength Index (RSI) is a momentum indicator that helps assess the strength of price trends, offering insights into potential reversals or continuations in stock movement.

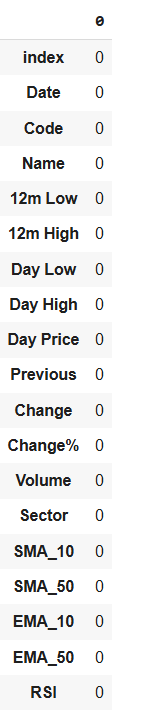
RSI is calculated based on average gains and losses over a specified period (typically 14 days). Readings above 70 suggest overbought conditions, meaning the stock might be due for a correction, while readings below 30 indicate oversold conditions, suggesting potential buying opportunities.

When price movements contradict RSI trends, it may signal trend reversals—for instance, if prices hit new highs while RSI forms lower highs, it could indicate weakening momentum.

A high RSI indicates that a security has been experiencing strong upward price movements, while a low RSI suggests strong downward price movements.

RSI is a powerful tool for gauging the internal strength of price movements, helping to identify potential turning points and providing valuable momentum context for robust financial analysis and predictive modeling.

This is how the data set looks with the new features.



* 1. Adding Lag Features.

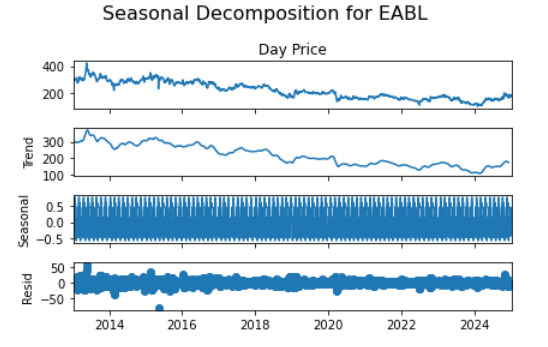
Lag features are a crucial technique in time-series analysis, where existing data points are shifted backward by one or more time steps to create new predictive variables. This approach enhances the dataset by introducing historical dependencies, allowing models to recognize trends, seasonality, and temporal relationships that influence future values. By transforming a simple time series into a structured tabular format, each row not only contains present observations but also relevant past data, enriching the input for supervised machine learning models.

Incorporating lag features significantly improves forecasting accuracy, as models can leverage past fluctuations and patterns to make more informed predictions. This method is particularly valuable in financial markets, demand forecasting, and anomaly detection, where understanding previous behaviors helps anticipate future outcomes.

# **CHAPTER 7. MODELING**

To perform forecasting, a combination of SARIMA, XGBoost, and LSTM models was employed, each contributing unique strengths to time series analysis. SARIMA (Seasonal AutoRegressive Integrated Moving Average) is an extension of the traditional ARIMA model that effectively captures both linear trends and seasonal patterns in time series data. XGBoost, a robust gradient boosting algorithm, improves predictive performance by modeling complex relationships and highlighting important features. LSTM (Long Short-Term Memory), a type of recurrent neural network, is well-suited for detecting long-term dependencies and nonlinear dynamics in sequential data, making it particularly useful for modeling the volatility of financial time series. This ensemble approach offers a well-rounded framework for forecasting, combining statistical structure, machine learning flexibility, and deep learning sophistication.

Checking for seasonality.



The repeated, tightly packed wave-like pattern suggests strong seasonality in the stock prices, likely influenced by periodic business cycles, holidays, or financial quarters.

The amplitude (height of oscillation) appears relatively consistent over time.

The seasonal decomposition of EABL Day Price reveals that the primary driver of its long-term movement is the Trend component, which showed a significant decline followed by stabilization and modest recovery. A consistent, yet relatively small-magnitude, seasonal pattern is also present, contributing to short-term predictability. Finally, the Residuals indicate that while much of the price movement is explained, there are still notable unpredictable events or noise influencingthe price, especially evident in the occasional large spikes.

**7.1 SARIMA (Seasonal AutoRegressive Integrated Moving Average)**

**Mean Absolute Error:**

An MAE of 3.3668 means that, on average, the model's predictions are off by about 3.37 units from the true values.

Since MAE is less sensitive to outliers, this suggests that most errors are relatively small.

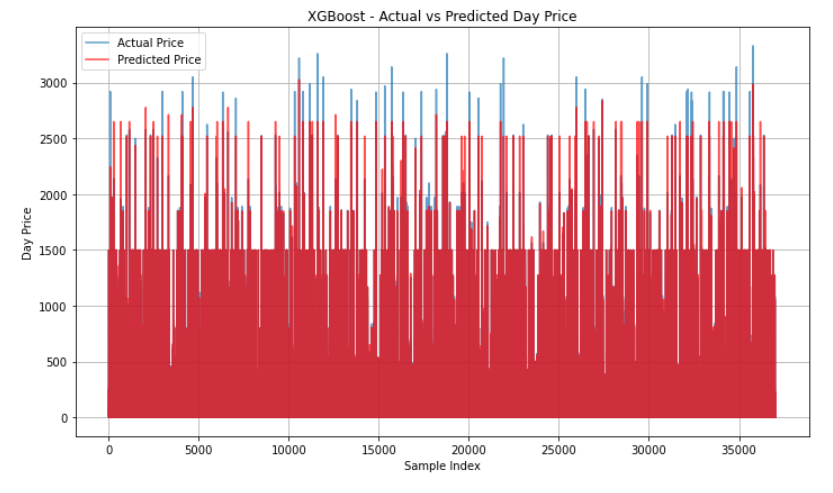
**Root Mean Squared Error (RMSE:**

An RMSE of **3.8853** means that the typical prediction error is around **3.89 units** when considering larger deviations.

Since RMSE > MAE, this indicates that some predictions have larger errors (outliers or occasional big misses).

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model demonstrates a reasonable level of accuracy. This indicates that, on average, its predictions are close to actual values. However, the slightly elevated Root Mean Square Error (RMSE) suggests that while the model performs well overall, it occasionally produces larger deviations from expected results. RMSE accounts for both minor and significant errors, meaning some predictions may have a higher margin of error due to volatility in the underlying data or seasonal fluctuations.

**7.2 XGBoost**



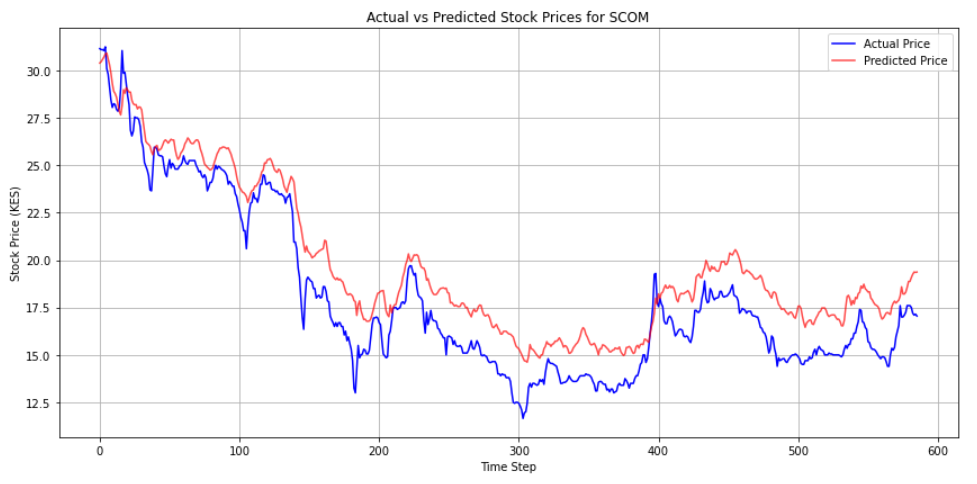
These performance metrics highlight the model’s exceptional predictive capability, making it highly reliable for forecasting applications. The **Mean Absolute Error (MAE) of 0.972** suggests that, on average, the model's predictions are very close to actual values, indicating minimal deviation. This is particularly valuable in cases where small errors can impact decision-making, ensuring the model provides precise estimates.

The **Root Mean Squared Error (RMSE) of 2.099** further emphasizes the model's accuracy. Since RMSE accounts for both minor and significant errors by squaring deviations before averaging them, it provides a more weighted measure of predictive quality. While RMSE is slightly higher than MAE due to this squared error effect, the relatively low value suggests that extreme errors are rare, reinforcing the model’s overall consistency.

Meanwhile, the **R² score of 0.979** showcases the model's ability to explain nearly all variability in the dataset. This high value indicates that the model captures essential patterns and trends, leaving only a small fraction of variance unaccounted for. A strong R² score is particularly useful when evaluating how well the model generalizes across different data points, ensuring it does not merely fit historical data but remains effective in predicting new observations.

Taken together, these metrics confirm that the model performs exceptionally well, maintaining low error rates while demonstrating robust predictive stability. If further refinements are needed, adjustments such as feature selection, hyperparameter tuning, or ensemble techniques could enhance accuracy even more. However, given the current evaluation, this model stands out as a highly effective tool for making reliable predictions with minimal risk of large deviations.

**7.3 LSTM (Long Short-Term Memory)**



The **Mean Absolute Error (MAE) of 3.9024** indicates that, on average, the model’s predictions deviate by approximately 3.9 units from actual values. Since MAE calculates the absolute difference between predicted and real values without considering direction, it serves as a straightforward measure of accuracy. Its insensitivity to outliers means that the majority of prediction errors are moderate rather than extreme, making it a useful metric for evaluating overall consistency.

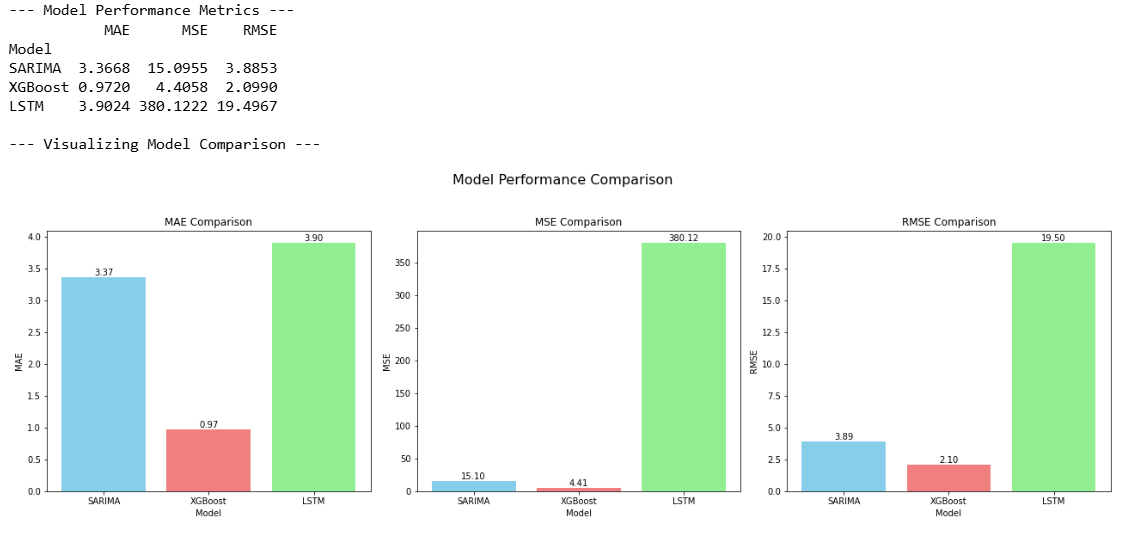
Meanwhile, the **Mean Squared Error (MSE) of 380.1222** highlights how the model penalizes larger errors more heavily. Because MSE squares each error before averaging, significant deviations exert a greater influence on the final value. This means that while small errors remain manageable, larger inaccuracies have a disproportionately high impact on the metric. While MSE itself can be difficult to interpret in absolute terms, its value is crucial for assessing variance in prediction errors and determining how well the model minimizes large deviations.

The **Root Mean Squared Error (RMSE) of 6.2345** provides a more intuitive and interpretable representation of typical error magnitude. As the square root of MSE, RMSE brings error values back to the original scale, making it easier to understand in the context of real-world data. A higher RMSE suggests that, while predictions are generally close to actual values, occasional larger deviations exist that contribute to overall error. Depending on the scale of the dataset, an RMSE of 6.2345 may be considered reasonable or significant.

Overall, these metrics collectively provide a balanced assessment of the model’s forecasting accuracy. While the relatively low MAE suggests stable errors, the higher MSE and RMSE indicate that some larger deviations may occasionally occur. Understanding these aspects allows for targeted improvements in model tuning, feature selection, or data preprocessing to further refine predictive precision.

# **CHAPTER 8. MODEL EVALUATION.**

This section details the performance evaluation of the candidate models – SARIMA, XGBoost, and LSTM – based on key regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The objective is to identify the model exhibiting the superior predictive capability for the task at hand.

****

The performance evaluation clearly highlights the superiority of the XGBoost model over both the SARIMA and LSTM models in terms of accuracy and error minimization. Across all three metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)—XGBoost consistently delivers the lowest error values, demonstrating its ability to capture complex data patterns efficiently. Its gradient boosting framework optimizes predictions iteratively, reducing deviations from actual values and ensuring strong generalization across different time series trends.

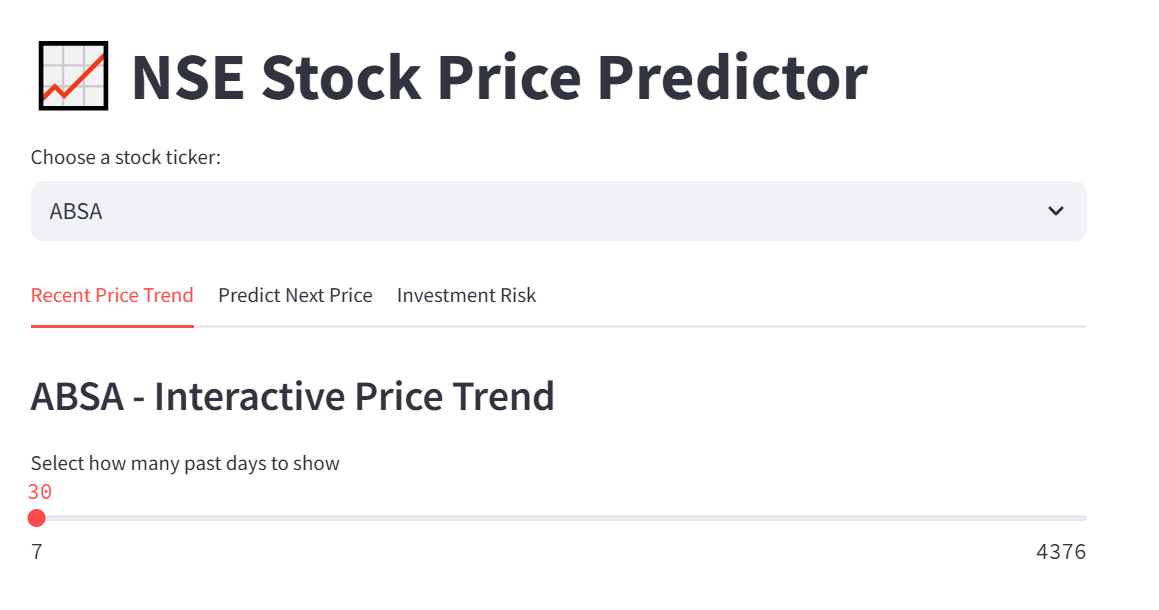
SARIMA, while performing better than LSTM, falls short compared to XGBoost. Its strength lies in capturing seasonal dependencies and structured temporal trends, making it suitable for time series forecasting. However, its limitation in adapting to nonlinear relationships and dynamic market movements results in higher errors compared to XGBoost, which can handle both structured and unstructured patterns more effectively.

The LSTM model, in contrast, exhibits significantly larger error values, particularly in MSE and RMSE, indicating greater variability in its predictions. While LSTM is designed for sequential data processing and excels in learning long-term dependencies, its performance in this evaluation suggests difficulties in stabilizing short-term fluctuations and adapting to the dataset's characteristics. This could be due to challenges in hyperparameter tuning, insufficient training epochs, or the inherent complexity of the dataset affecting its learning efficiency. Additionally, recurrent models like LSTM may struggle with sudden shifts in trends if not optimized correctly, leading to higher predictive errors.

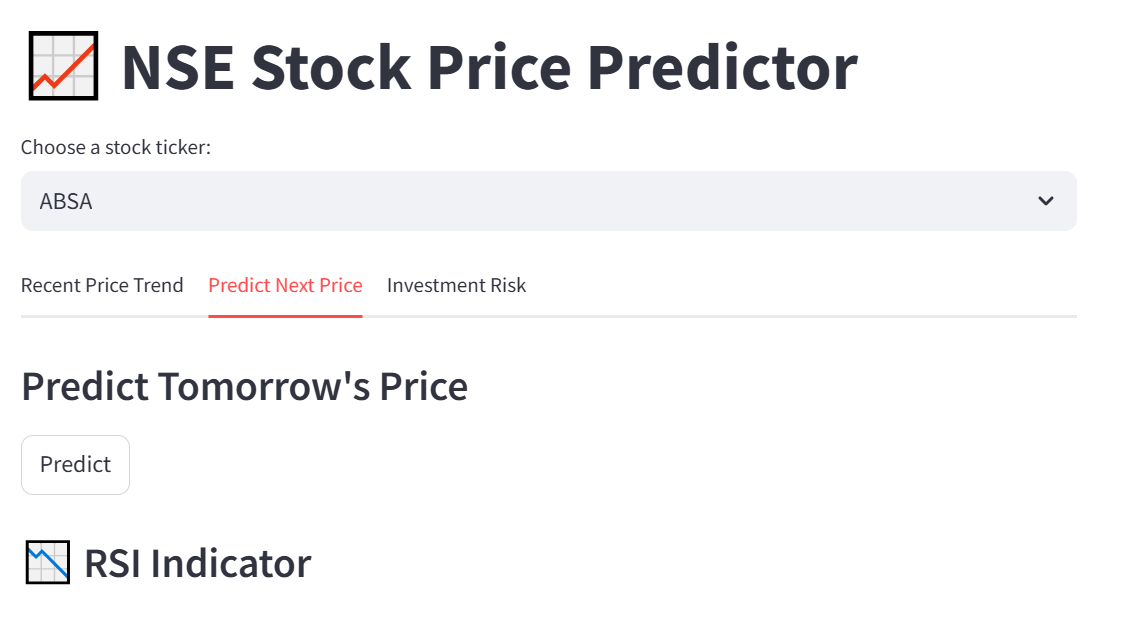
Overall, the results reaffirm XGBoost as the most reliable model for this particular forecasting task, given its ability to minimize errors while maintaining robust predictive stability. While SARIMA remains a reasonable alternative for traditional time series applications, LSTM's weaker performance suggests that further optimization and adjustments may be required to enhance its accuracy in similar evaluations.

# **CHAPTER 9. DEPLOYMENT**

We deployed the XGBoost model using streamlit. These are some of the features we explored.



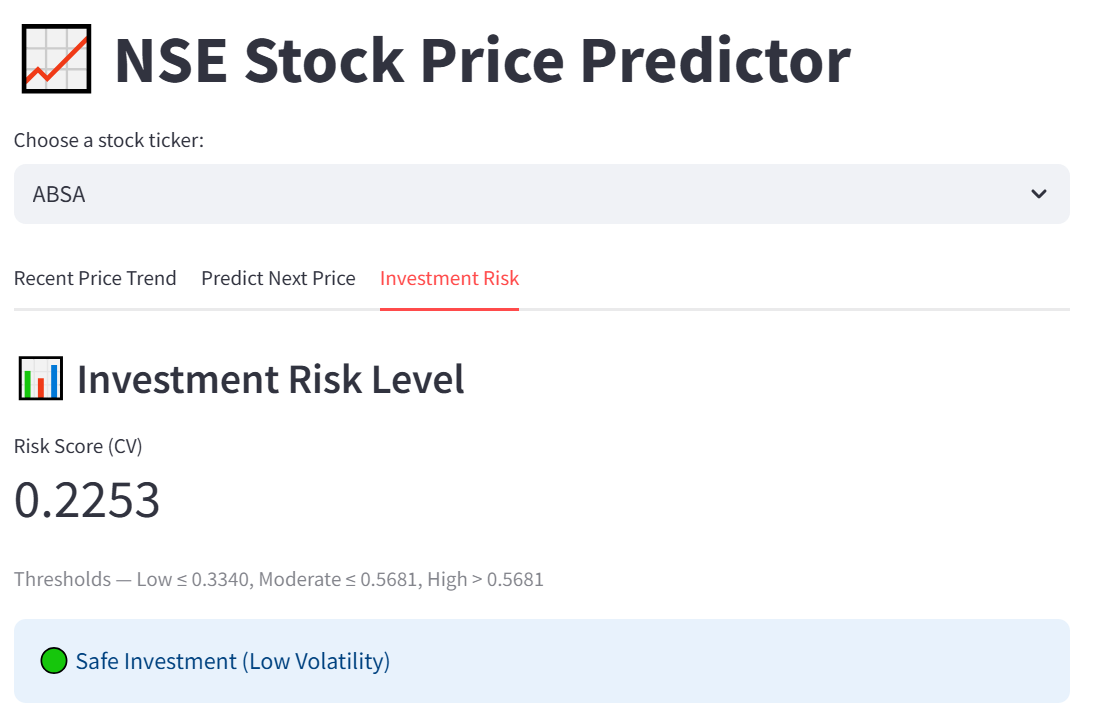
It allows users to select a stock ticker (currently showing ABSA) and provides options to view the recent price trend, predict the next price, and assess investment risk. The visible section of the page is focused on displaying the "ABSA - Interactive Price Trend", with a slider to adjust the number of past days shown, currently set to 30.



In this view, the "Predict Next Price" option is selected. The main section is dedicated to predicting tomorrow's price for the chosen stock, which is currently ABSA, and includes a "Predict" button.

This tab allows the user after selecting the stock they wish to predict. It provides the prediction and also further information is provided using the RSI indicator.

The RSI indicator indicates if the stock is over bought or oversold.



This is the "Investment Risk" section of the "NSE Stock Price Predictor" web application. It displays the investment risk level for the selected stock, which is ABSA. The page shows a "Risk Score (CV)" of 0.2253 and provides thresholds to interpret this score: Low (≤0.3340), Moderate (≤0.5681), and High (>0.5681). Based on the calculated score, the application determines that ABSA is currently a "Safe Investment (Low Volatility)".

# **CHAPTER 10. CONCLUSIONS AND RECOMMENDATIONS.**

**10.1 CONCLUSIONS**

1. The XGBoost model clearly outperforms both the SARIMA and LSTM models. It consistently records the lowest error values across all three measures, indicating superior predictive accuracy and robustness. This suggests that XGBoost is particularly effective at capturing complex patterns in the data.
2. Effectiveness in Handling Volatility and Non-linearity: Stock market data, including the NSE, is inherently volatile and exhibits complex non-linear relationships. The strong performance of XGBoost indicates its effectiveness in modeling these characteristics. As a tree-based ensemble method, XGBoost is well-suited to capture complex interactions between features and handle non-linear patterns that simpler linear models like ARIMA may struggle with.
3. While LSTM is designed to handle sequential data and long-term dependencies effectively, its performance may have been impacted by challenges in adapting to the dataset’s structure. Additionally, recurrent neural networks like LSTM can sometimes struggle with short-term fluctuations in the NSE data which it faced potentially leading to larger errors in its predictive tasks.

**10.2 RECOMMENDATIONS**

1. Focus on high-volume stocks, as they typically provide better liquidity, allowing traders to enter and exit positions with minimal price slippage. High trading volume also indicates strong investor interest and activity, which often leads to more consistent and significant price movements. These characteristics make high-volume stocks particularly attractive for short-term trading strategies, where timely execution and frequent price action are essential for capturing quick gains.
2. Consider allocating a larger share of liquidity-sensitive investments to the Telecommunication and Banking sectors, as these typically exhibit higher trading volumes and market activity. This enhanced liquidity facilitates easier entry and exit from positions, reducing the likelihood of price slippage and potentially lowering transaction costs. Additionally, these sectors often include well-established companies with steady investor interest, making them more stable and predictable for managing large or frequently adjusted positions. This makes them ideal for strategies that require flexibility and quick response to market conditions.
3. Diversify your portfolio by including both high-volume sectors and emerging industries to achieve a balanced risk-return profile. High-volume sectors, such as Finance and Telecommunications, offer stability, liquidity, and consistent performance due to their established market presence and strong investor participation. In contrast, emerging sectors—like Technology, Renewable Energy, or Fintech—may exhibit higher volatility but present significant growth potential as they innovate and expand. This diversification strategy helps cushion against market downturns in any single sector while positioning the portfolio to benefit from future growth opportunities in rapidly evolving industries.

**10.3 NEXT STEPS**

1. Enhance Feature Engineering and Integration by:

Adding technical indicators like Bollinger Bands to enrich model input.

Incorporating volume-based metrics to reflect market liquidity.

Integrating fundamental and macroeconomic data like earnings reports and sector performance.

2. Expanding Data Utilization and Continuous Retraining by:

Defining an optimal historical data window for modeling.

Automating data ingestion and model retraining through pipelines.

Exploring high-frequency data (hourly/intraday) for improved short-term forecasting.

3. Improving Model Evaluation and Performance Monitoring by:

Using advanced metrics like Sharpe Ratio (risk-adjusted return),Maximum Drawdown (worst peak-to-trough loss) and Hit Rate (accuracy of directional predictions).

Implementing drift detection to identify and react to market regime changes.