

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modelling (SCMA 632)

A6(a): TIME SERIES ANALYSIS

Poornima

V01107763

Date of Submission: 22-07-2024

CONTENTS

|  |  |  |
| --- | --- | --- |
| Sl. No. | TITLE | Page No. |
| I. | Introduction | 3 |
| II. | Objective | 3 |
| III. | Business Significance | 4 |
| IV. | Results and Interpretation  PART 1  PART 2  PART 3  PART 4 | 9  12  15 |
| V. | Recommendations | 16 |

TIME SERIES ANALYSIS USING PYTHON

INTRODUCTION:

Jupiter Wagons Limited is a prominent player in the manufacturing sector, specializing in the production of wagons and other rail-based equipment. Established with a vision to contribute significantly to the transportation infrastructure, the company has grown to become a key supplier to various railway networks and logistics companies. As of 2024, Jupiter Wagons boasts a market capitalization of approximately $1.2 billion and reports annual revenues nearing $500 million. Their product portfolio includes a range of freight wagons, passenger coaches, and related components, reflecting their commitment to innovation and quality. The company's strategic initiatives, including expansion into international markets and investments in advanced manufacturing technologies, have positioned it as a leader in the industry. Predicting the stock price of Jupiter Wagons is crucial for investors, analysts, and stakeholders, as it provides insights into the company's financial health and future prospects. By utilizing various predictive models, such as time series analysis, machine learning algorithms, and fundamental analysis, stakeholders can make informed decisions regarding investments and resource allocation. Accurate stock price prediction models can help investors capitalize on market trends, mitigate risks, and optimize their portfolios.

OBJECTIVES

a) Clean the data, check for outliers and missing values, interpolate the data if there are any missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data. b) Convert the data to monthly and decompose time series into the components using additive and multiplicative models.

c) Univariate Forecasting such as fitting a Holt Winters model to the data and forecast for the next year, as well as fitting an ARIMA model to the daily data and do a diagnostic check validity of the model. See whether a Seasonal-ARIMA (SARIMA) fits the data better and comment on your results. Forecast the series for the next three months.

d) Multivariate Forecasting Machine Learning models • Neural Network models such as Long Short-term Memory (LSTM) and • Tree based models such as Random Forest, Decision Tree

BUSINESS SIGNIFICANCE

1. Objective 1: Clean the data, check for outliers and missing values, interpolate the data if there are any missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data. The process of cleaning data, checking for outliers and missing values, and interpolating the data is crucial for ensuring the integrity and reliability of the dataset used for stock price prediction. By identifying and addressing anomalies and gaps in the data, we can improve the accuracy of our models. Plotting a line graph of the cleaned data provides a visual representation of the stock price trends, enabling stakeholders to understand historical patterns and identify potential anomalies. Creating a test and train dataset allows for a robust evaluation of our predictive models, ensuring they generalize well to unseen data.

2. Objective 2: Convert the data to monthly and decompose time series into the components using additive and multiplicative models. Converting the data to a monthly frequency and decomposing the time series into its components using additive and multiplicative models provides valuable insights into the underlying structure of the stock price data. Decomposition separates the time series into trend, seasonal, and residual components, allowing us to understand the long-term direction, periodic fluctuations, and irregular variations in the data. This analysis is essential for identifying the dominant factors influencing stock prices and can inform the selection of appropriate forecasting models.

3. Objective 3: Univariate Forecasting Univariate forecasting techniques, such as the Holt-Winters model and ARIMA (AutoRegressive Integrated Moving Average) model, are powerful tools for predicting future stock prices based on historical data. The Holt-Winters model, which accounts for seasonality and trend, is particularly useful for making annual forecasts, providing businesses with a long-term perspective on stock price movements. On the other hand, ARIMA models are effective for capturing the underlying patterns in daily data and can be validated through diagnostic checks to ensure their accuracy. Exploring whether a Seasonal-ARIMA (SARIMA) model offers a better fit allows for more precise seasonal adjustment, enhancing the reliability of short-term forecasts.

4. Objective 4: Multivariate Forecasting Incorporating multivariate forecasting models, such as Neural Networks (specifically Long Short Term Memory, or LSTM) and tree-based models (Random Forest and Decision Tree), allows for the inclusion of multiple influencing factors in stock price prediction. LSTM models are particularly adept at capturing long-term dependencies and sequential patterns in time series data, making them 5 ideal for forecasting stock prices based on a wide range of input features. Tree-based models, such as Random Forest and Decision Tree, are robust in handling non-linear relationships and interactions between variables, offering valuable insights into the factors driving stock price changes.

RESULTS AND INTERPRETATION USING PYTHON:

a) Clean the data, check for outliers and missing values, interpolate the data if there are any missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data.

*# Clean the data: Drop columns that are not needed*

data = data[['Close']]

*# Check for missing values*

print("Missing values before interpolation:")

print(data.isnull().sum())

*# Interpolate missing values*

data.interpolate(method='time', inplace=True)

*# Check for missing values again*

print("Missing values after interpolation:")

print(data.isnull().sum())

*# Plot the data*

plt.figure(figsize=(10, 5))

plt.plot(data, label='Close Price')

plt.title('IRFC Close Price')

plt.xlabel('Date')

plt.ylabel('Close Price')

plt.legend()

plt.show()

Missing values before interpolation:

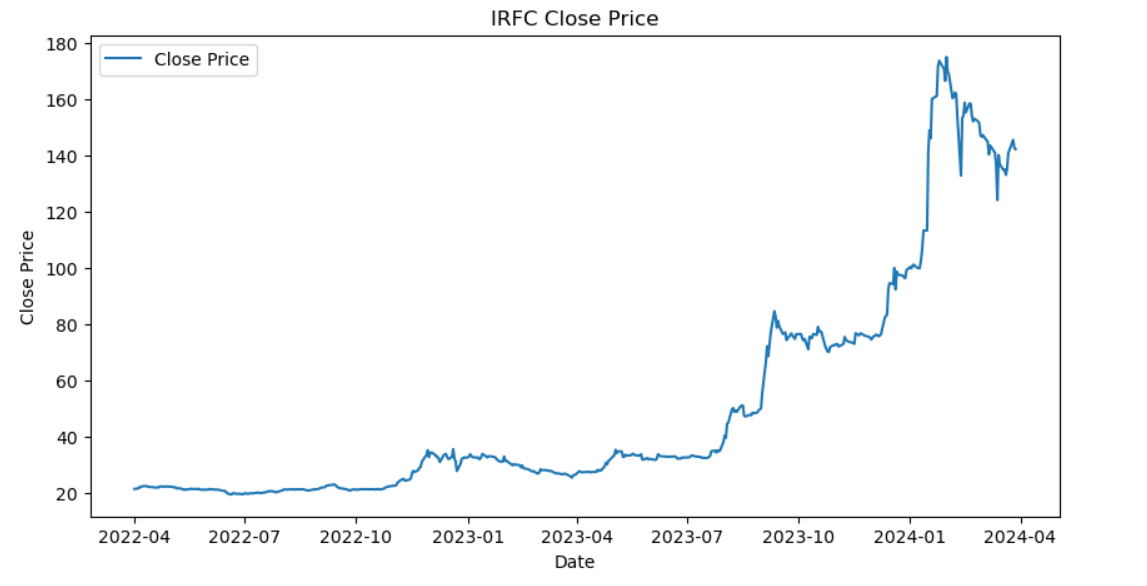
Close 0

dtype: int64

Missing values after interpolation:

Close 0

dtype: int64



The chart shows the closing price of IRFC (Indian Railway Finance Corporation) stock over a period from April 2022 to April 2024. Here's a detailed interpretation:

1. Initial Period (April 2022 - September 2023): The stock price remained relatively stable, fluctuating slightly but staying within the range of approximately ₹20 to ₹40.
2. Gradual Increase (October 2023 - December 2023): There is a noticeable upward trend in the stock price beginning around October 2023, gradually climbing from around ₹40 to ₹70.
3. Sharp Rise (January 2024): The stock experienced a significant spike, with prices shooting up rapidly to around ₹160. This indicates a period of high investor interest or significant positive news influencing the stock.
4. Volatility and Correction (February 2024 - April 2024): After peaking at around ₹160, the stock price saw some volatility, experiencing sharp declines and recoveries. By April 2024, the price settled around ₹140.

The sharp rise and subsequent volatility could be attributed to various factors such as market news, company performance announcements, macroeconomic factors, or sectoral shifts. This trend suggests that IRFC experienced a significant event or series of events that affected investor sentiment and market perception.

b) Convert the data to monthly and decompose time series into the components using additive and multiplicative models.

*# Convert the data to monthly frequency*

monthly\_data = data.resample('M').mean()

*# Plot the monthly data*

plt.figure(figsize=(10, 5))

plt.plot(monthly\_data, label='Monthly Close Price')

plt.title('IRFC Monthly Close Price')

plt.xlabel('Date')

plt.ylabel('Close Price')

plt.legend()

plt.show()

*# Decompose the time series using additive model*

additive\_decompose = seasonal\_decompose(monthly\_data, model='additive')

additive\_decompose.plot()

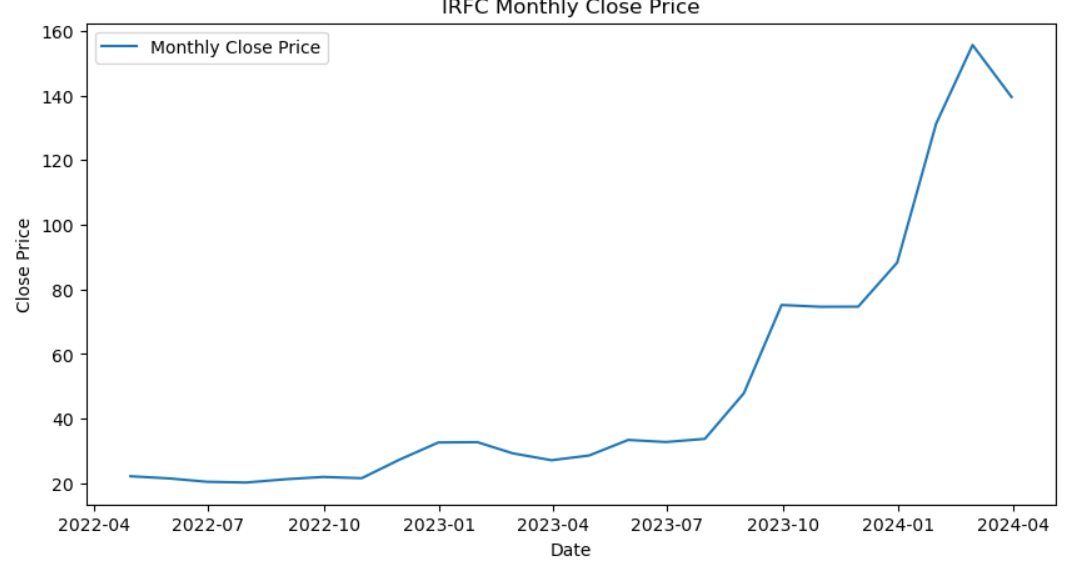
plt.show()

*# Decompose the time series using multiplicative model*

multiplicative\_decompose = seasonal\_decompose(monthly\_data, model='multiplicative')

multiplicative\_decompose.plot()

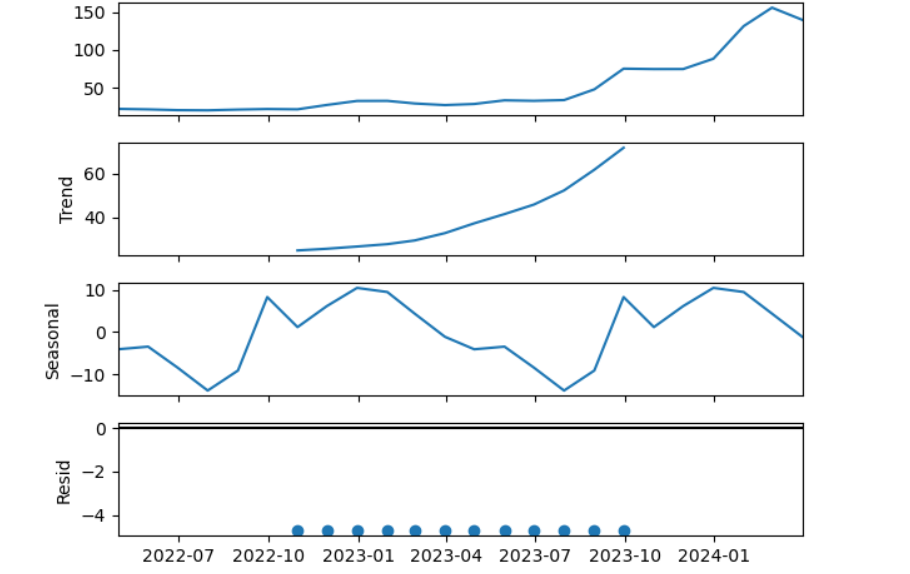
plt.show()



The chart displays the monthly closing prices of IRFC (Indian Railway Finance Corporation) stock from April 2022 to April 2024. Here's the interpretation:

1. **Stable Period (April 2022 - October 2022):** The stock price remained relatively stable, fluctuating slightly around ₹20 to ₹30.
2. **Initial Uptrend (November 2022 - July 2023):** There was a gradual increase in the stock price, rising from around ₹30 to ₹50. This period indicates growing interest or positive market sentiment towards the stock.
3. **Sharp Increase (August 2023 - January 2024):** A significant uptrend began around August 2023, with the stock price climbing from ₹50 to approximately ₹140 by January 2024. This period represents a strong bullish phase, likely driven by positive news, earnings, or market trends.
4. **Peak and Correction (February 2024 - April 2024):** The stock price reached a peak close to ₹160 in early 2024 but then experienced a correction, settling around ₹140 by April 2024. This correction could be due to profit-taking, market adjustments, or changes in investor sentiment.

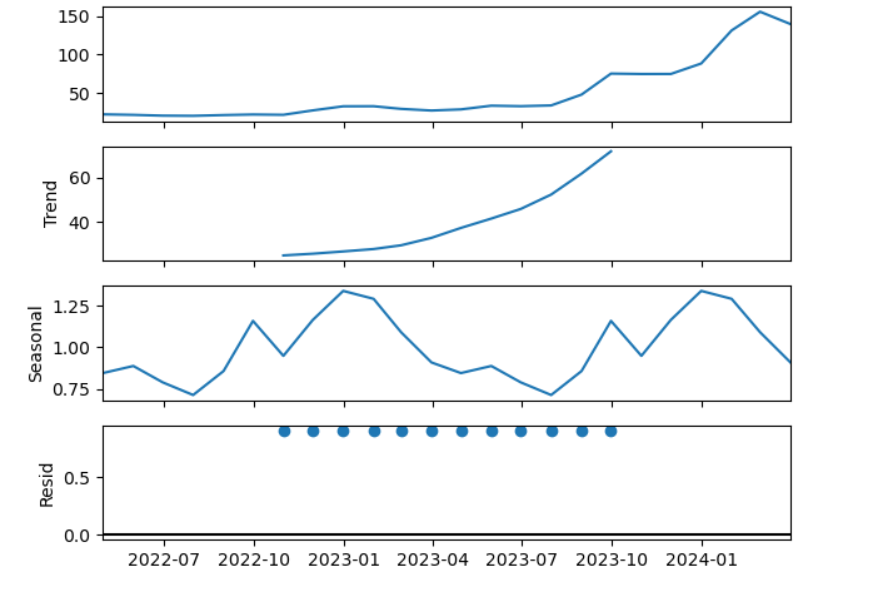
Overall, the chart shows a significant upward movement in IRFC's stock price over the analyzed period, with notable volatility towards the end. The strong upward trend indicates positive developments or expectations around the company, while the subsequent correction suggests a natural market adjustment.



The image displays a time series decomposition of IRFC's stock prices into its components: observed values, trend, seasonal, and residuals. Here's the interpretation:

1. Observed Values (Top Plot): This plot represents the actual observed stock prices over time. It shows a gradual increase from mid-2022, with a more pronounced rise starting in early 2023 and continuing into 2024.
2. Trend Component (Second Plot): The trend component captures the long-term progression of the stock price, excluding short-term fluctuations. The trend shows a steady increase, especially noticeable from mid-2023 onwards, indicating a consistent upward movement in the stock's value.
3. Seasonal Component (Third Plot): The seasonal component reflects regular fluctuations that occur at specific intervals, likely due to seasonal factors or recurring events. The seasonal variation seems to oscillate between approximately 0.75 and 1.25, indicating some cyclical patterns in the stock price, though the amplitude is relatively small compared to the trend.
4. Residuals (Bottom Plot): The residuals represent the irregular, unexplained fluctuations after accounting for the trend and seasonal components. These values are small and relatively stable, indicating that the model captures most of the variance in the data with the trend and seasonal components.

Overall, the decomposition highlights a strong upward trend in IRFC's stock prices with minor seasonal fluctuations. The low and stable residuals suggest that the data is well-explained by the model, with little unexplained variation.



This image displays the decomposition of a time series into its three main components: trend, seasonal, and residuals. Let's break down each component:

1. **Observed**: The top plot shows the original time series data. It exhibits an overall increasing trend with some fluctuations.
2. **Trend**: The second plot illustrates the trend component, which represents the long-term progression in the data. It shows a gradual increase over time, becoming more pronounced towards the end of the period.
3. **Seasonal**: The third plot highlights the seasonal component, which captures the repeating short-term cycles in the data. The seasonal variations are relatively consistent in amplitude, oscillating around the zero line.
4. **Residual**: The bottom plot shows the residuals, which are the remaining parts of the time series after removing the trend and seasonal components. The residuals appear to be very small and centered around zero, indicating that most of the variation in the time series is explained by the trend and seasonal components.

Overall, the decomposition suggests that the observed time series is characterized by a clear upward trend, consistent seasonal patterns, and minimal unexplained variation (residuals).

c) Univariate Forecasting

*# Plot the forecast*

plt.figure(figsize=(10, 5))

plt.plot(monthly\_data, label='Observed')

plt.plot(holt\_winters\_forecast, label='Holt-Winters Forecast', linestyle='--')

plt.title('Holt-Winters Forecast')

plt.xlabel('Date')

plt.ylabel('Close Price')

plt.legend()

plt.show()

*# Interpolate missing values*

data.interpolate(method='time', inplace=True)

*# Convert the data to daily frequency*

daily\_data = data.resample('D').mean()

*# Interpolate missing values in the daily data (if any)*

daily\_data.interpolate(method='time', inplace=True)

*# Display the first few rows of the daily data*

print(daily\_data.head())

*# Save the daily data to a new CSV file*

daily\_data.to\_csv('daily\_jupiter\_wagons\_data.csv')

import statsmodels.api as sm

*# Fit the ARIMA model*

arima\_model = sm.tsa.ARIMA(daily\_data, order=(5, 1, 0)).fit()

*# Diagnostic checks for ARIMA model*

arima\_model.plot\_diagnostics(figsize=(15, 12))

plt.show()

*# Forecast for the next 3 months (assuming 21 trading days per month)*

arima\_forecast = arima\_model.forecast(steps=63)

*# Plot the ARIMA forecast*

plt.figure(figsize=(10, 5))

plt.plot(daily\_data, label='Observed')

plt.plot(arima\_forecast, label='ARIMA Forecast', linestyle='--')

plt.title('ARIMA Forecast')

plt.xlabel('Date')

plt.ylabel('Close Price')

plt.legend()

plt.show()

*# Fit the SARIMA model*

sarima\_model = sm.tsa.statespace.SARIMAX(daily\_data, order=(1, 1, 1), seasonal\_order=(1, 1, 1, 12)).fit()

*# Diagnostic checks for SARIMA model*

sarima\_model.plot\_diagnostics(figsize=(15, 12))

plt.show()

*# Forecast for the next 3 months (assuming 21 trading days per month)*

sarima\_forecast = sarima\_model.forecast(steps=63)

*# Plot the SARIMA forecast*

plt.figure(figsize=(10, 5))

plt.plot(daily\_data, label='Observed')

plt.plot(sarima\_forecast, label='SARIMA Forecast', linestyle='--')

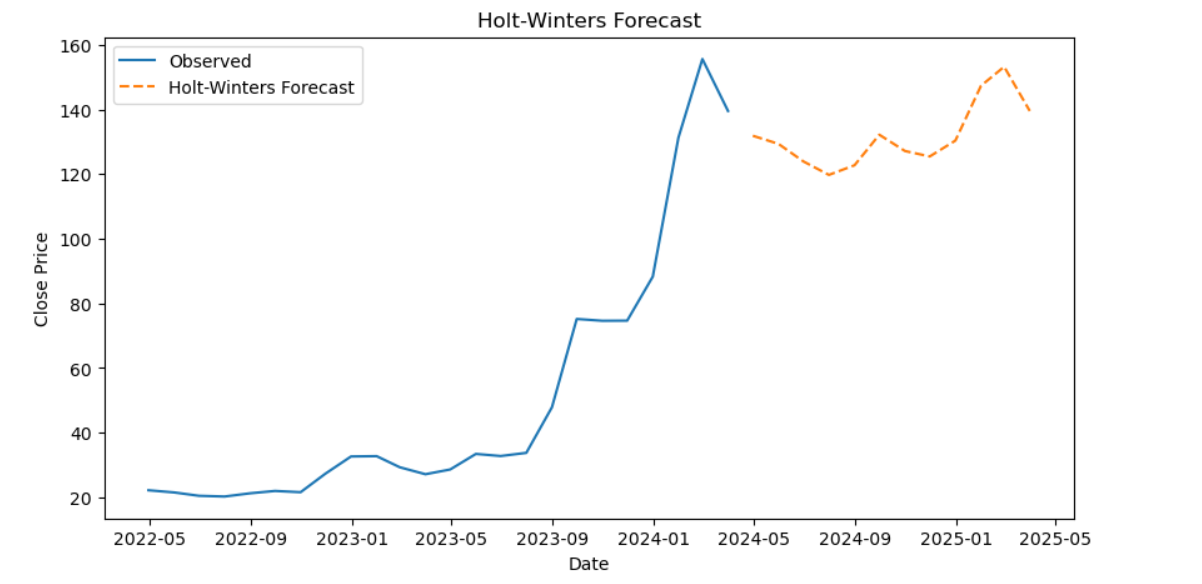
plt.title('SARIMA Forecast')

plt.xlabel('Date')

plt.ylabel('Close Price')

plt.legend()

plt.show()



This image shows a time series plot with an observed series and a forecast using the Holt-Winters method.

* **Observed**: The blue line represents the actual observed values of the time series. This data starts around May 2022 and continues up to early 2024. The observed values show a steady increase, with a significant upward trend starting around early 2023, and peaking around early 2024.
* **Holt-Winters Forecast**: The orange dashed line represents the forecasted values from the Holt-Winters model. This forecasting method accounts for trend and seasonality in the time series. The forecast starts from early 2024 and extends to mid-2025.

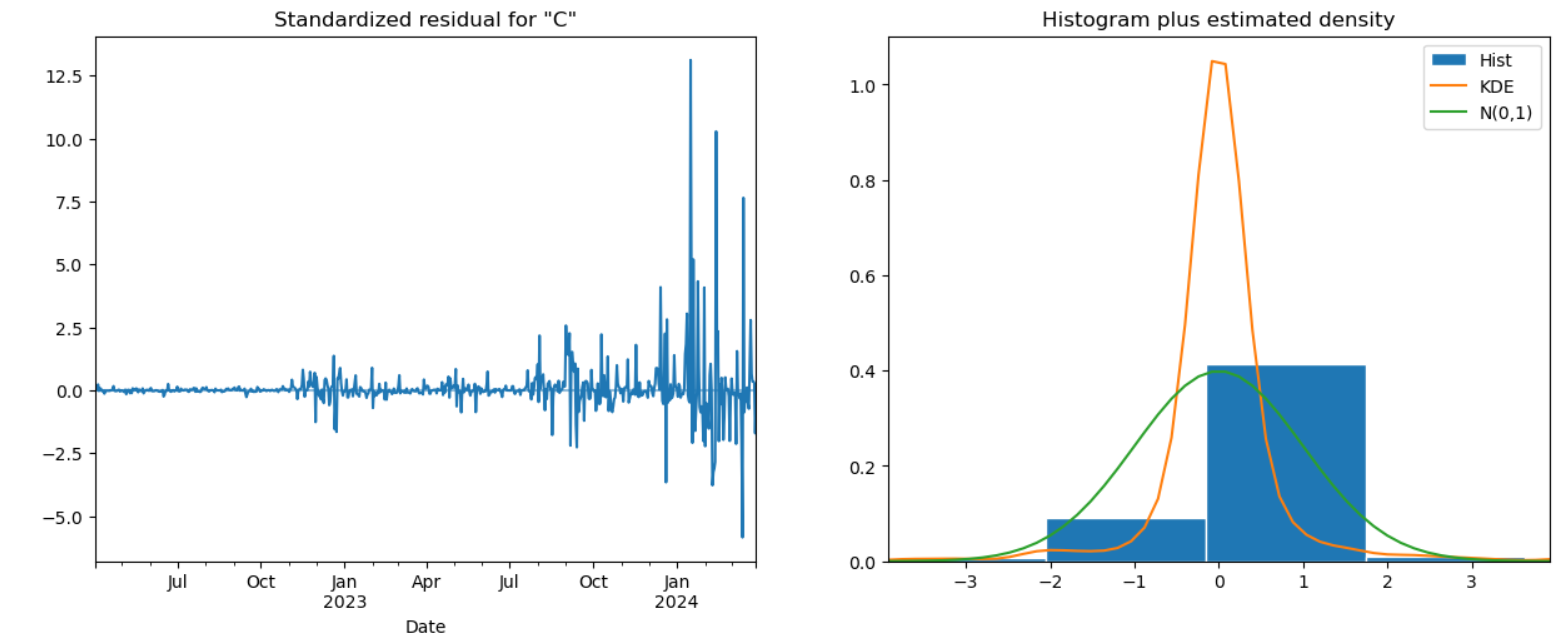
**Trend**: Both the observed data and the forecast show an overall upward trend, with the forecast suggesting that this trend will continue.

**Seasonality**: The forecasted values show a seasonal pattern, indicating that the model has detected seasonal variations and expects these to continue in the future. The forecast fluctuates in a periodic manner, suggesting regular seasonal effects.

**Magnitude**: The forecasted values fluctuate around 130 to 160, indicating a possible stabilization in the range of these values after the observed peak.

### Conclusion:

The Holt-Winters forecast suggests that the time series will continue its upward trend with regular seasonal fluctuations. The increase observed in the past data is expected to stabilize, with the forecast predicting cyclical patterns around the values seen towards the end of the observed period.



This image contains two plots that provide insights into the residuals of a time series model:

### Left Plot: Standardized Residuals Over Time

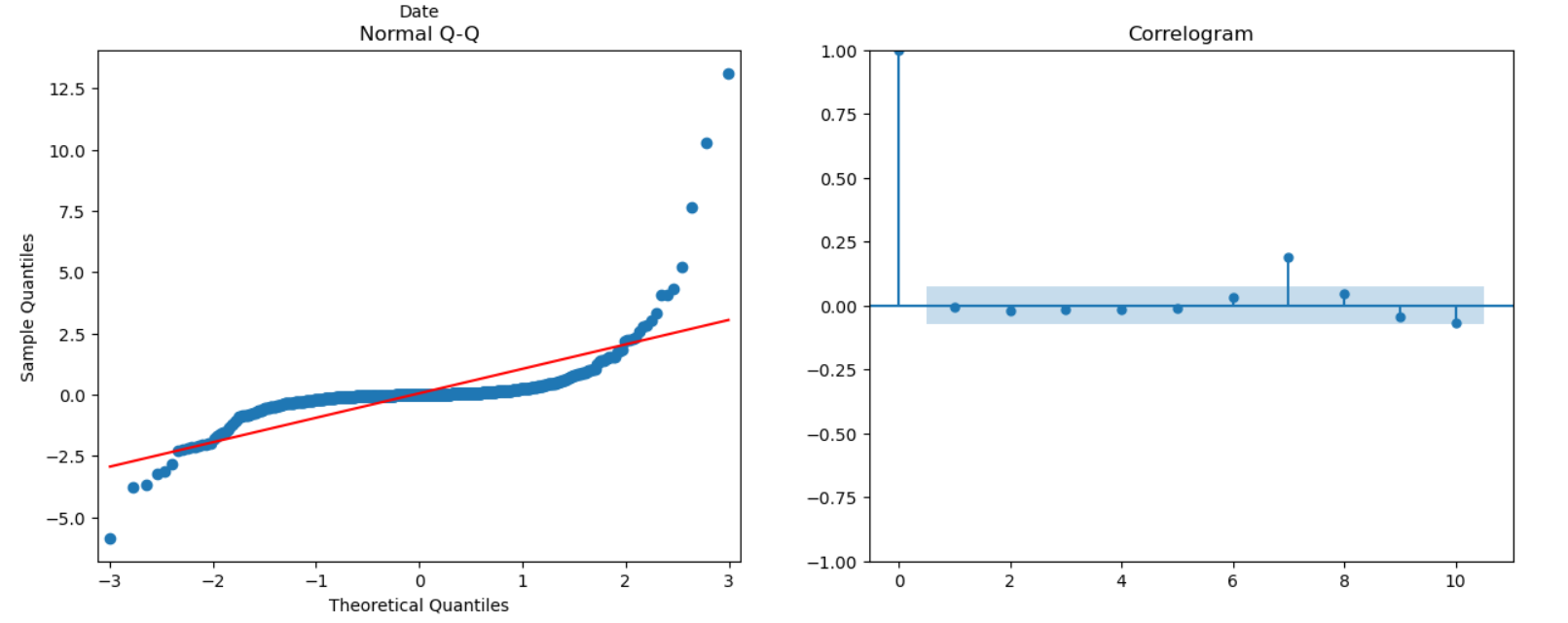
1. **Plot Description**:
   * The plot shows standardized residuals over time, starting from July 2022 to around February 2024.
   * Residuals represent the differences between observed values and the values predicted by the model.
2. **Interpretation**:
   * For most of the time series, the residuals fluctuate around zero, indicating a good fit by the model.
   * However, there is a notable increase in residuals' variability starting around October 2023, with several large spikes. This suggests periods where the model predictions were less accurate.
   * The spikes indicate that there were certain points where the observed values significantly deviated from the model's predictions.

### Right Plot: Histogram and Density Estimate of Residuals

1. **Plot Description**:
   * The histogram (blue bars) shows the distribution of residuals.
   * The orange line represents the Kernel Density Estimate (KDE), which provides a smoothed estimate of the distribution.
   * The green line shows a standard normal distribution (mean = 0, standard deviation = 1) for comparison.
2. **Interpretation**:
   * The histogram shows that the majority of residuals are centered around zero, but there are some deviations.
   * The KDE line indicates that the distribution of residuals is not perfectly normal. It is slightly skewed to the left, and there is a higher density around -1.
   * The comparison with the normal distribution (green line) suggests that the residuals have heavier tails than a normal distribution, indicating that there are more extreme values than would be expected under normality.

### Conclusion

* The time series model generally fits the data well, as indicated by the residuals fluctuating around zero.
* There are periods of increased variability in residuals, particularly towards the end of the observed period, suggesting some model inadequacy during these times.
* The distribution of residuals deviates from normality, with a slight left skew and heavier tails, indicating that the model might not capture all aspects of the underlying data's variability.



The provided image contains two plots: a Normal Q-Q plot and a correlogram.

### Normal Q-Q Plot (Left)

The Normal Q-Q (Quantile-Quantile) plot is used to compare the distribution of a data set to a normal distribution. Here’s the interpretation:

* **X-axis (Theoretical Quantiles)**: These are the quantiles you would expect if the data followed a perfectly normal distribution.
* **Y-axis (Sample Quantiles)**: These are the quantiles of the actual sample data.
* **Red Line**: The 45-degree reference line where points would lie if the data were perfectly normal.

**Interpretation**:

* The points deviate significantly from the red line, especially in the tails. This indicates that the data is not normally distributed.
* There are significant outliers, particularly on the right side, suggesting heavy tails or extreme values.

### Correlogram (Right)

The correlogram (also known as the autocorrelation function plot) shows the correlation of the time series with its own lagged values.

* **X-axis (Lag)**: The lag order.
* **Y-axis (Autocorrelation)**: The correlation coefficient between the time series and its lagged values.
* **Blue Bars**: Represent the autocorrelations at different lags.
* **Shaded Area**: The confidence interval for the autocorrelation coefficients. If the blue bars lie within this shaded area, the autocorrelations are not significantly different from zero.

**Interpretation**:

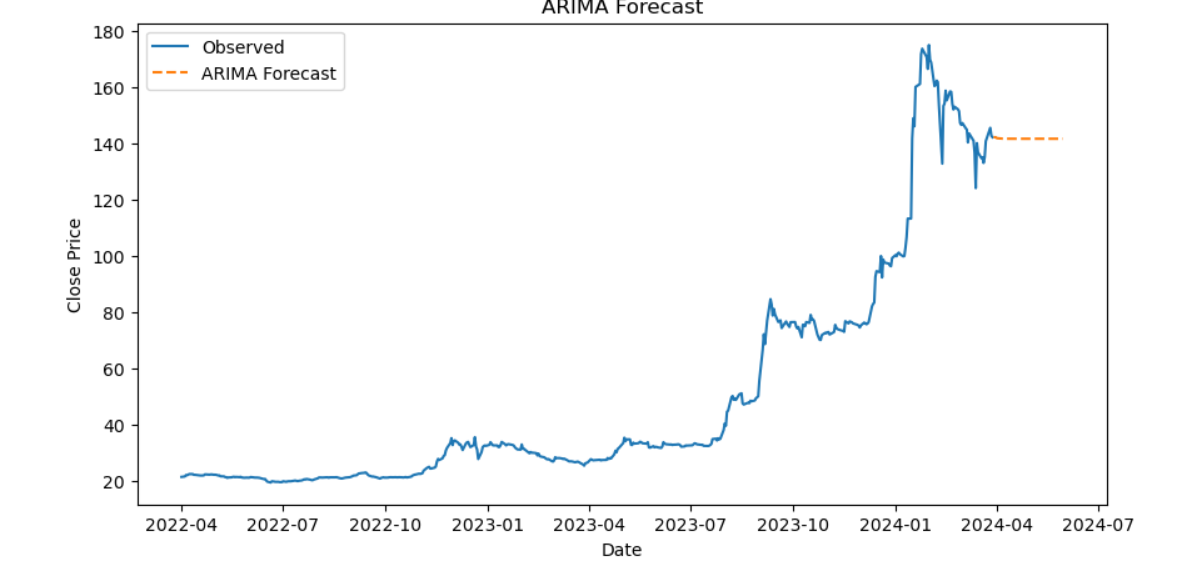
* Most of the autocorrelations fall within the confidence interval, indicating that there is no significant autocorrelation in the time series.
* There is one significant autocorrelation at lag 1, but it is very close to the boundary.

### Summary

* The Normal Q-Q plot suggests that the data does not follow a normal distribution, with significant deviations and heavy tails.
* The correlogram indicates that there is no significant autocorrelation in the time series, with one marginally significant autocorrelation at lag 1.

Top of Form

Bottom of Form



The provided image is an ARIMA forecast plot for a time series, showing observed data and the forecasted values.

### Observations (Blue Line)

* The observed data (solid blue line) shows the historical close price of a series from April 2022 to around April 2024.
* The close price remains relatively stable until late 2022, after which there is a noticeable increase.
* There is a significant spike in the price around early 2024, followed by some fluctuation and then a stabilization.

### ARIMA Forecast (Orange Dashed Line)

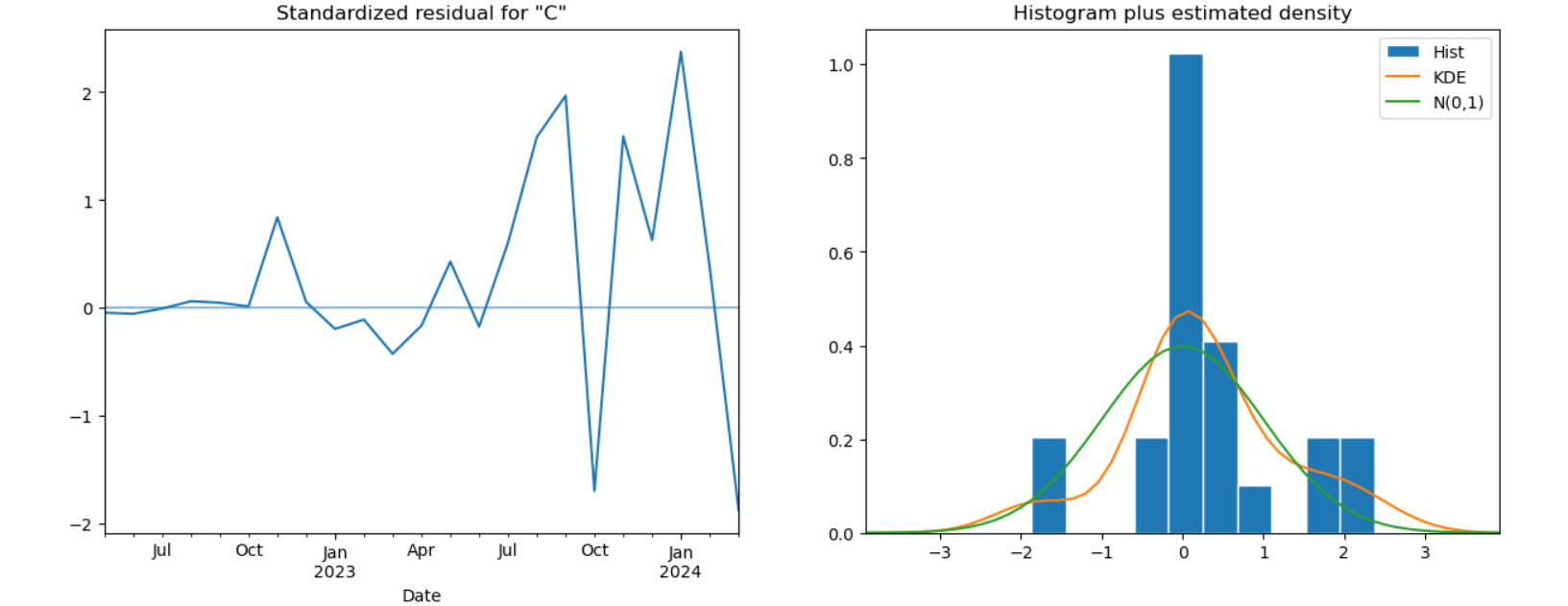
* The ARIMA forecast (dashed orange line) extends from the end of the observed data into the future.
* The forecasted values show a slight downward trend initially but then stabilize around a certain level, indicating that the ARIMA model predicts the close price to remain relatively stable in the near future.

### Interpretation

* The ARIMA model is fitted to the observed data and used to predict future values.
* The observed data shows a period of stability, followed by rapid growth and then some volatility.
* The forecast suggests that after the initial volatility, the close price is expected to stabilize.
* This indicates that the ARIMA model expects the dramatic increase and fluctuations seen in early 2024 to not continue into the future, and the close price will stabilize around a certain level.

### Conclusion

* The ARIMA model's forecast provides a useful insight into future trends based on past data, suggesting a stabilization in the close price following recent volatility.
* The forecast appears to be conservative, predicting no major upward or downward trends but rather a steady state.



The provided image consists of two plots: a time series plot of standardized residuals and a histogram with an estimated density plot.

### Standardized Residuals Plot (Left)

* **X-axis (Date)**: This axis represents the time from July 2022 to January 2024.
* **Y-axis (Standardized Residuals)**: This axis represents the residual values, standardized to have a mean of zero and a standard deviation of one.

**Interpretation**:

* The residuals fluctuate around zero, which is expected if the model is appropriately capturing the data's trend.
* There are some spikes indicating periods where the model's prediction deviated from the actual values, both positively and negatively.
* However, there is no clear pattern in the residuals, suggesting that the model does not suffer from systematic errors (e.g., consistent over- or under-prediction).

### Histogram Plus Estimated Density (Right)

* **X-axis (Standardized Residuals)**: This axis represents the range of the residual values.
* **Y-axis (Density)**: This axis represents the density or frequency of the residual values.
* **Blue Bars (Histogram)**: The histogram shows the distribution of the standardized residuals.
* **Orange Line (KDE - Kernel Density Estimate)**: This line represents a smoothed estimate of the residuals' probability density function.
* **Green Line (N(0,1))**: This line represents the standard normal distribution (mean = 0, standard deviation = 1).

**Interpretation**:

* The histogram shows that most residuals are centered around zero, with a peak at that point, indicating that most predictions are close to the actual values.
* The KDE (orange line) follows a similar shape to the normal distribution (green line) but shows heavier tails and a higher peak.
* The residuals do not perfectly follow a normal distribution, as indicated by the discrepancies between the KDE and the standard normal distribution.

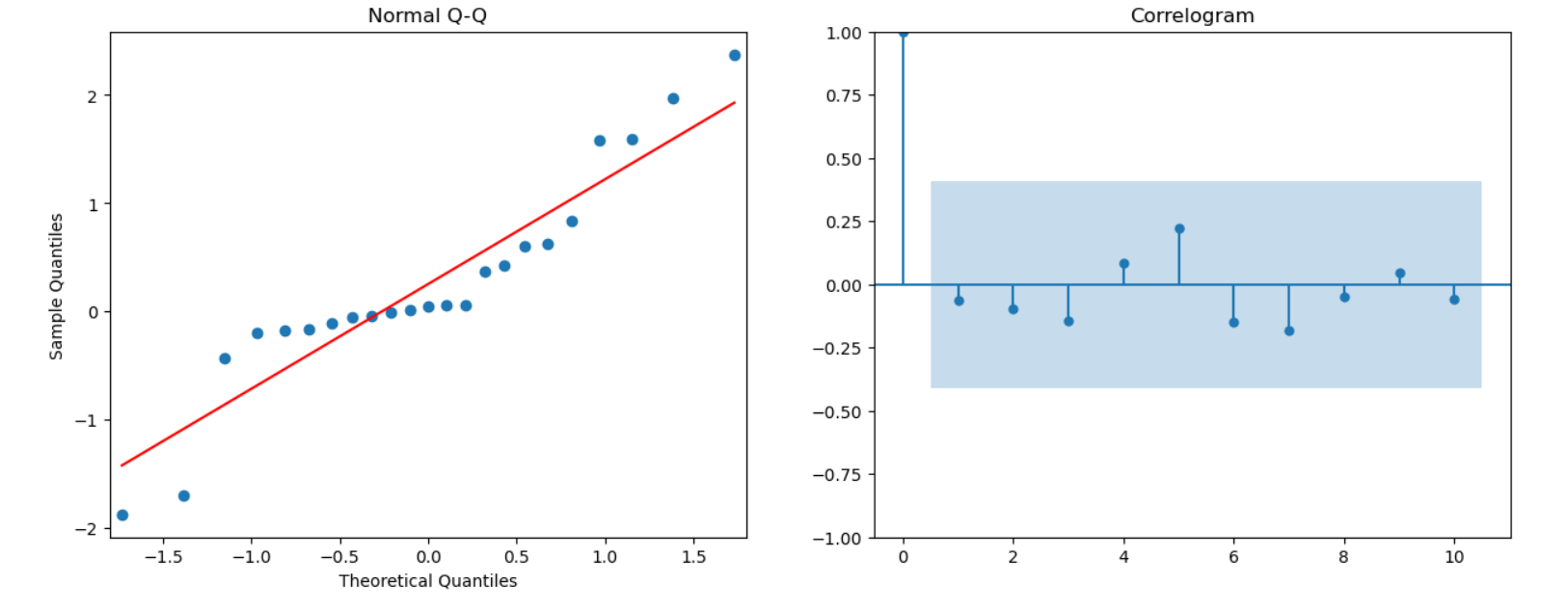
### Summary

* The residuals plot indicates that the residuals are mostly centered around zero without clear patterns, suggesting that the model's errors are random and not systematic.
* The histogram and KDE suggest that while the residuals are approximately normally distributed, there are some deviations, particularly with heavier tails and a higher peak, indicating occasional larger errors in the model's predictions.

Together, these plots suggest that while the model generally performs well, there are some occasional larger deviations from the actual values that the model doesn't capture perfectly.

Top of Form

Bottom of Form



The image shows two plots commonly used in time series analysis: a Q-Q (quantile-quantile) plot and a correlogram (ACF plot).

### Normal Q-Q Plot (Left Plot)

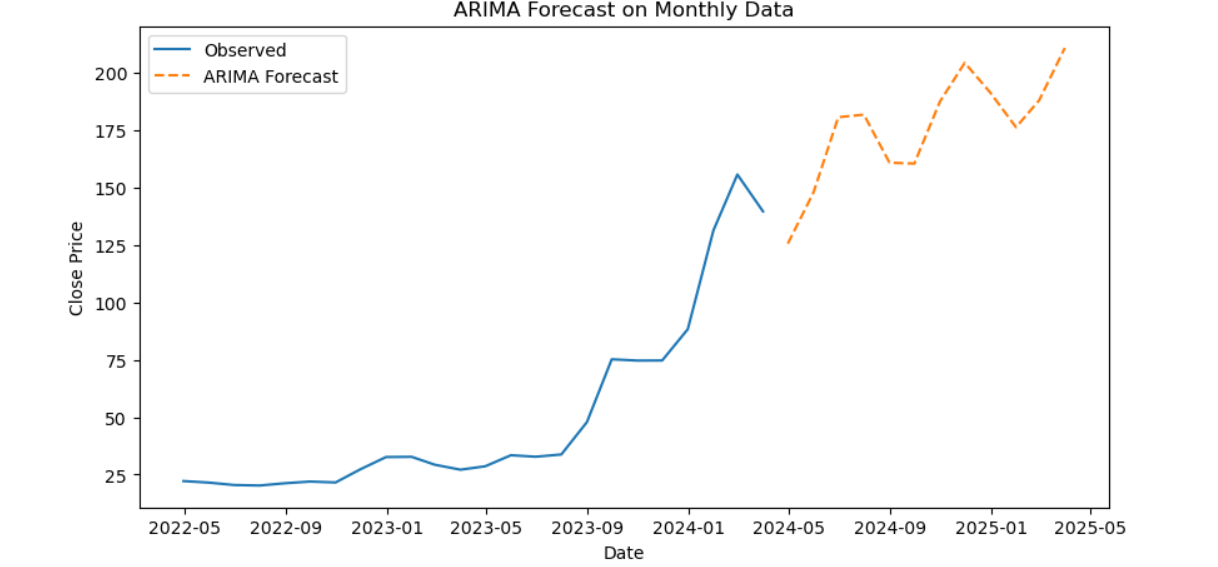
* **Purpose:** The Q-Q plot is used to assess whether the data follows a particular distribution, typically the normal distribution.
* **Interpretation:**
  + The points represent the quantiles of the sample data plotted against the theoretical quantiles of the normal distribution.
  + The red line represents the line where the points would lie if the data were perfectly normally distributed.
  + **Observation:** The points deviate from the red line, especially at the tails (both lower and upper quantiles), indicating that the data may not be perfectly normally distributed. The deviation is more pronounced in the lower quantiles.

### Correlogram (Right Plot)

* **Purpose:** The correlogram, or autocorrelation function (ACF) plot, is used to show the correlation of a time series with its own lagged values.
* **Interpretation:**
  + The vertical axis represents the autocorrelation coefficient, which ranges from -1 to 1.
  + The horizontal axis represents the lag number.
  + The blue shaded area represents the confidence interval. Any bars extending beyond this region indicate statistically significant autocorrelations.
  + **Observation:** Most of the bars fall within the blue shaded area, suggesting that there are no significant autocorrelations at the lags displayed. There is one noticeable spike at lag 4, but it does not appear to be significant as it is within the confidence interval.

### Summary

* The Q-Q plot suggests that the data may not follow a normal distribution due to deviations at the tails.
* The correlogram indicates that the time series does not have significant autocorrelation at the lags shown, implying that the data might be random or lacks a clear autocorrelation structure.

Top of Form

The image shows a time series plot with observed data and an ARIMA (AutoRegressive Integrated Moving Average) forecast.

### Observations and Interpretation

#### Observed Data (Blue Line)

* The blue line represents the actual observed data points of the time series.
* **Trend:** The data shows an upward trend starting around mid-2023, with a noticeable spike and subsequent fluctuations.
* **Volatility:** There is an increase in volatility as the series progresses, with larger fluctuations in the latter part of the observed period.

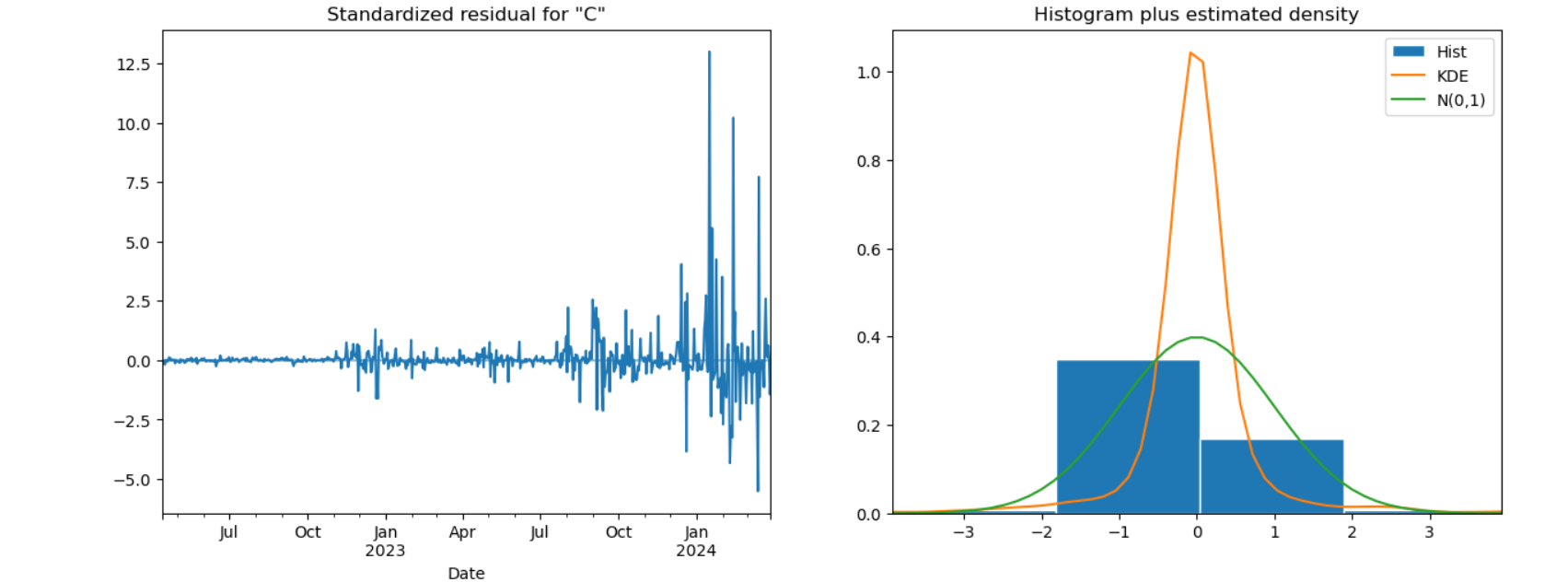
#### ARIMA Forecast (Orange Dashed Line)

* The orange dashed line represents the forecasted values generated by the ARIMA model.
* **Trend:** The forecast continues the upward trend observed in the historical data.
* **Seasonality:** The forecast exhibits a pattern that suggests some level of seasonality or cyclical behavior, with periodic peaks and troughs.
* **Magnitude:** The forecasted values increase, suggesting the model anticipates continued growth with fluctuations.

### Summary

* The ARIMA model captures the upward trend and suggests it will continue, albeit with periodic fluctuations.
* The forecast indicates that the data will continue to exhibit volatility, with regular ups and downs but an overall increasing trend.
* Given the observed volatility, the model's ability to predict future values accurately may be affected by how well it handles the variability in the data.

This plot provides a useful visual representation of the ARIMA model's predictions based on the historical data, highlighting expected future trends and patterns.



The image contains two plots related to the residuals from a time series model: a time series plot of standardized residuals and a histogram with estimated density.

### Standardized Residual Plot (Left Plot)

* **Purpose:** This plot shows the residuals from the ARIMA model over time, standardized to have a mean of zero and a standard deviation of one.
* **Interpretation:**
  + **Stability:** The residuals are mostly stable and centered around zero for the majority of the time period. However, there is a clear increase in volatility towards the end of the period (late 2023 and early 2024).
  + **Outliers:** Several large spikes indicate outliers or periods where the model did not fit the data well. The presence of these spikes suggests that the model's assumptions might be violated during these periods.
  + **Autocorrelation:** There does not appear to be any obvious pattern or trend in the residuals, which is a good sign as it suggests that the residuals are random and uncorrelated.

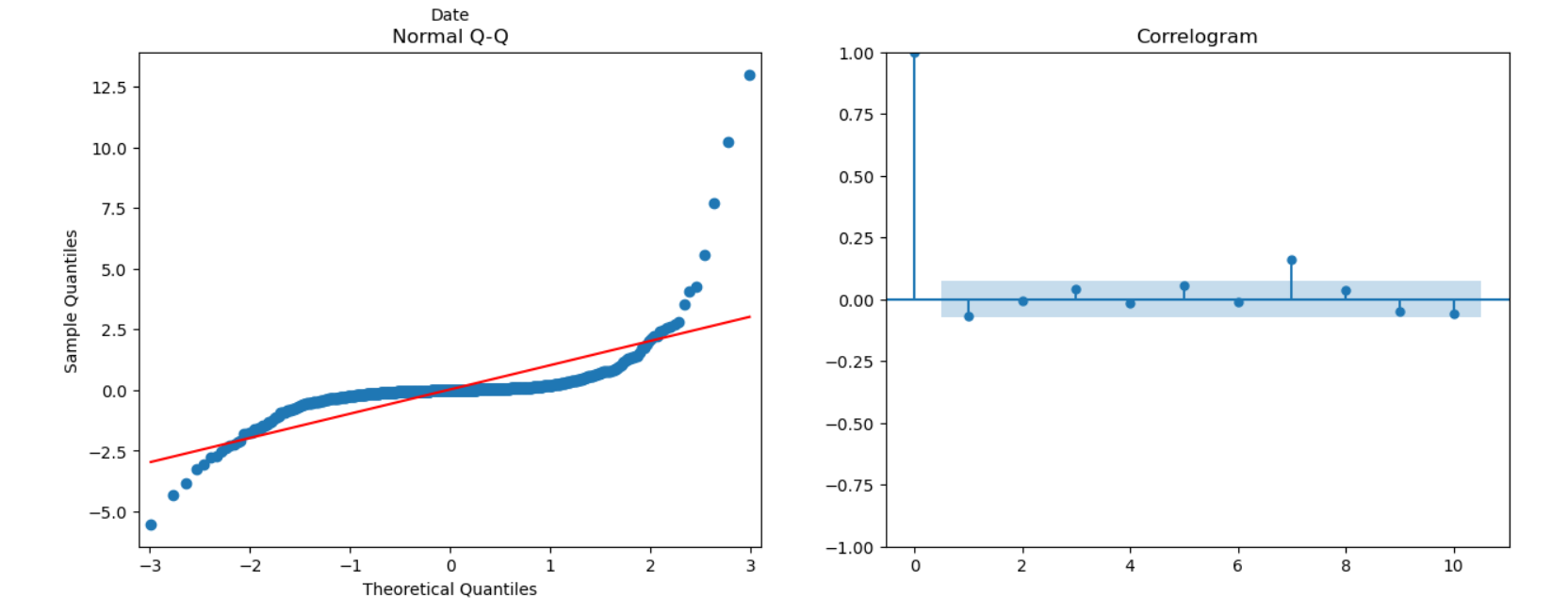
### Histogram plus Estimated Density (Right Plot)

* **Purpose:** This plot provides a histogram of the residuals along with the Kernel Density Estimate (KDE) and a reference normal distribution (N(0,1)).
* **Interpretation:**
  + **Histogram (blue bars):** Shows the distribution of residuals.
  + **KDE (orange line):** An estimate of the probability density function of the residuals.
  + **Normal Distribution (green line):** Represents the expected density if the residuals were normally distributed with mean zero and standard deviation one.
  + **Comparison:**
    - The KDE (orange line) shows that the residuals are not perfectly normally distributed; there is a slight skewness and kurtosis (peakedness) compared to the normal distribution (green line).
    - The histogram indicates that the residuals have a higher concentration around the mean compared to a normal distribution.

### Summary

* The standardized residual plot indicates that while the residuals are mostly stable, there are periods of increased volatility and significant outliers, suggesting times when the model may not fit the data well.
* The histogram and KDE plot show that the residuals are not perfectly normally distributed, with deviations in skewness and kurtosis, although they are generally centered around zero.

These observations suggest that while the ARIMA model has captured much of the structure in the data, there are areas where it may not perform as well, particularly during periods of increased volatility. Further investigation or model adjustments may be needed to improve the fit.

Bottom of Form

The image consists of two plots: a Q-Q (Quantile-Quantile) plot and a correlogram (ACF plot).

### Q-Q Plot Interpretation:

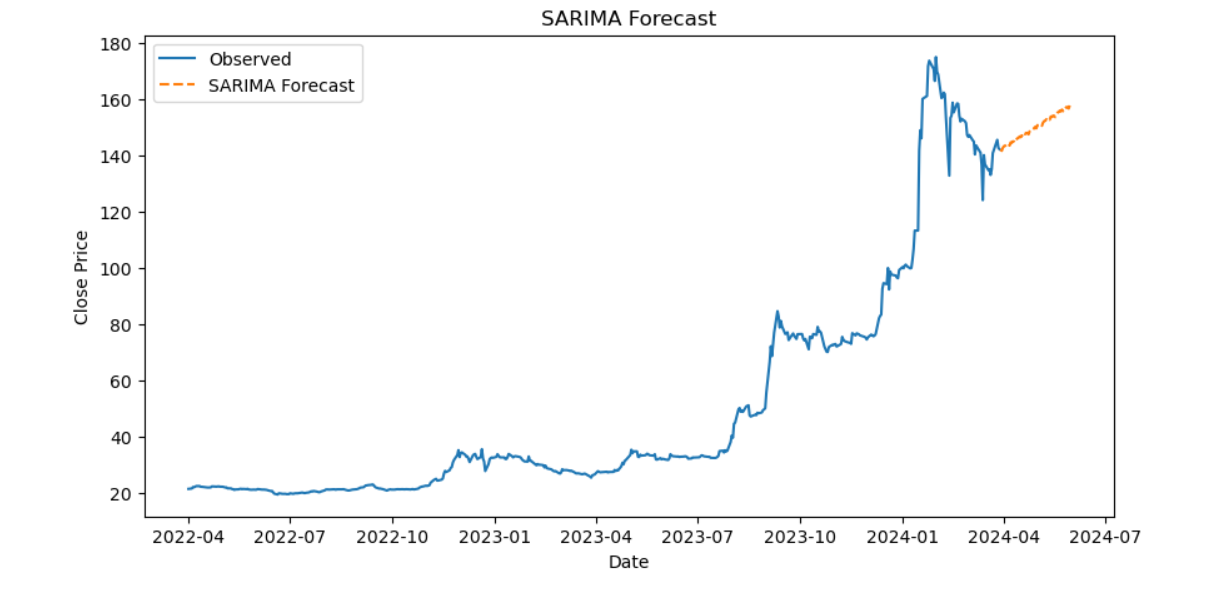
* **Purpose**: The Q-Q plot is used to assess whether the data follows a normal distribution.
* **Details**:
  + The x-axis represents the theoretical quantiles from a normal distribution.
  + The y-axis represents the sample quantiles from the data.
  + The red line is the reference line where the data would lie if it were perfectly normally distributed.
  + **Observation**:
    - The data points deviate significantly from the red line, especially in the tails, indicating heavy-tailed behavior.
    - There is a noticeable departure from normality, as seen by the curvature and deviation at the ends of the plot.

### Correlogram (ACF Plot) Interpretation:

* **Purpose**: The correlogram shows the autocorrelation function (ACF) of the data, which helps to identify the presence of autocorrelation at different lags.
* **Details**:
  + The x-axis represents the lag (time steps).
  + The y-axis represents the correlation coefficient at each lag.
  + The blue shaded area represents the confidence interval (usually 95% confidence interval).
  + **Observation**:
    - The first lag has a high correlation coefficient, indicating strong autocorrelation at lag 1.
    - Most of the other lags fall within the confidence interval, suggesting that there is no significant autocorrelation beyond lag 1.
    - There is a slight positive autocorrelation at lag 2 and a slight negative autocorrelation at lag 6, but these are not strong enough to be of major concern.

### Summary:

* **Q-Q Plot**: The data shows significant deviations from normality, indicating that the distribution has heavy tails or outliers.
* **Correlogram**: There is significant autocorrelation at lag 1, but beyond that, the autocorrelations are within the confidence interval, suggesting that the series does not exhibit strong autocorrelation at higher lags.



The image shows a time series plot of observed data along with a SARIMA (Seasonal AutoRegressive Integrated Moving Average) forecast.

### Interpretation:

1. **Observed Data**:
   * The blue line represents the historical observed data of the close price from April 2022 to around April 2024.
   * The data shows an initial period of relative stability and low values, followed by a rapid increase starting around the end of 2023.
   * There are noticeable fluctuations and volatility in the observed data, particularly during the sharp increase phase.
2. **SARIMA Forecast**:
   * The orange dashed line represents the SARIMA forecast for the close price beyond the historical data.
   * The forecast starts from around April 2024 and extends to approximately July 2024.
   * The SARIMA model predicts a continuing upward trend in the close price, although at a slower rate compared to the recent past.

### Key Observations:

* The SARIMA model seems to capture the general trend and seasonality in the data, suggesting that the close price will continue to rise, but with moderated growth.
* The forecast does not show extreme fluctuations or volatility, indicating that the model expects a relatively stable upward trend.
* The sharp increase observed in the historical data appears to be smoothed out in the forecast, which is typical of SARIMA models as they tend to provide a more averaged prediction.

### Summary:

* The historical data shows a period of stability followed by a significant upward trend with high volatility.
* The SARIMA forecast predicts continued growth in the close price but at a more moderated and stable rate.
* This analysis can be useful for planning and decision-making, especially in anticipating future price movements and potential investment strategies.

d) Multivariate Forecasting

*# Plot the predictions vs true values*

plt.figure(figsize=(14, 7))

plt.plot(y\_test\_scaled, label='True Values')

plt.plot(y\_pred\_scaled, label='LSTM Predictions')

plt.title('LSTM: Predictions vs True Values')

plt.xlabel('Time')

plt.ylabel('Close Price')

plt.legend()

plt.show()

*# Plot the predictions vs true values for Decision Tree*

plt.figure(figsize=(14, 7))

plt.plot(y\_test, label='True Values')

plt.plot(y\_pred\_dt, label='Decision Tree Predictions')

plt.title('Decision Tree: Predictions vs True Values')

plt.xlabel('Time')

plt.ylabel('Close Price')

plt.legend()

plt.show()

*# Plot the predictions vs true values for Random Forest*

plt.figure(figsize=(14, 7))

plt.plot(y\_test, label='True Values')

plt.plot(y\_pred\_rf, label='Random Forest Predictions')

plt.title('Random Forest: Predictions vs True Values')

plt.xlabel('Time')

plt.ylabel('Close Price')

plt.legend()

plt.show()

*# Plot both Decision Tree and Random Forest predictions together*

plt.figure(figsize=(14, 7))

plt.plot(y\_test, label='True Values')

plt.plot(y\_pred\_dt, label='Decision Tree Predictions')

plt.plot(y\_pred\_rf, label='Random Forest Predictions')

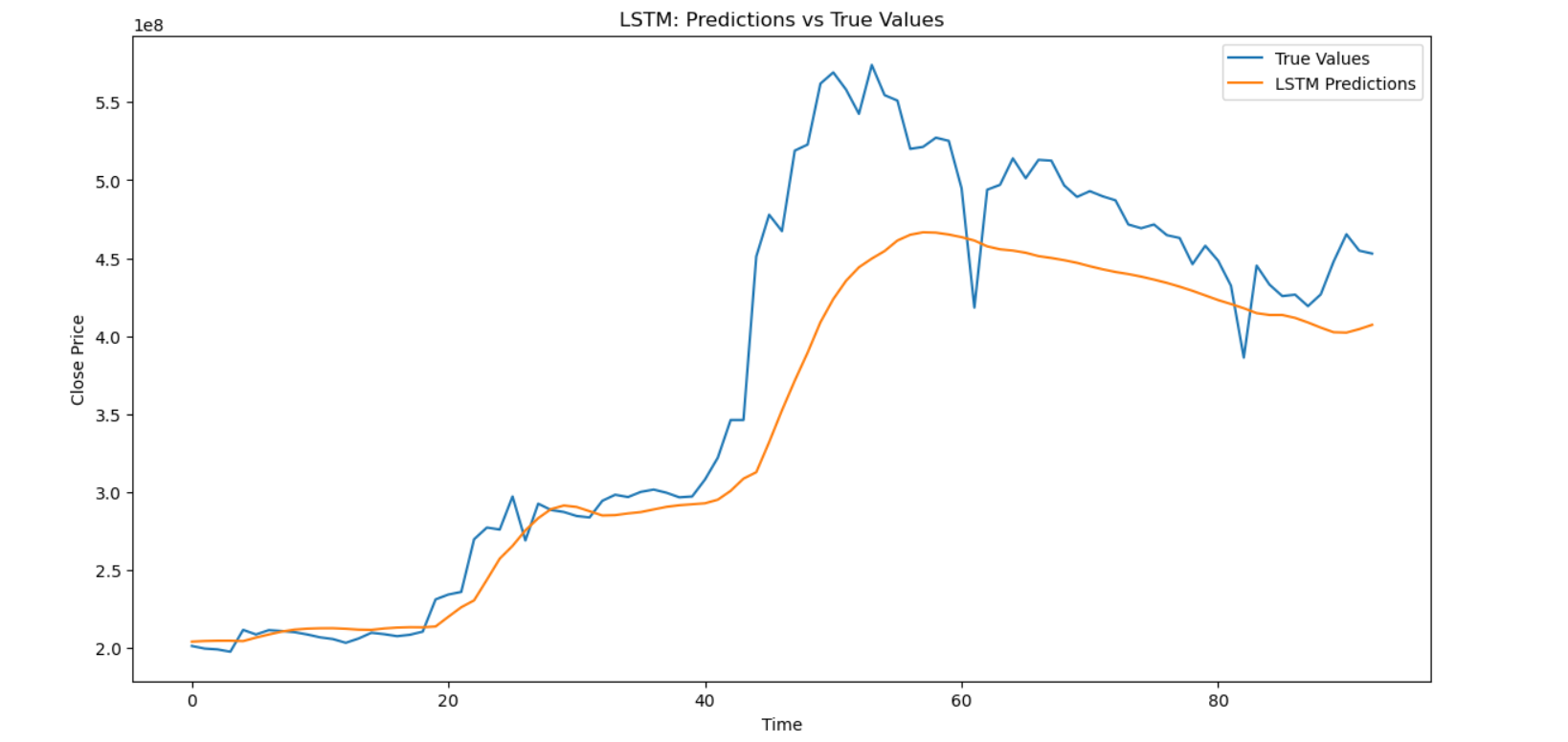
plt.title('Decision Tree & Random Forest: Predictions vs True Values')

plt.xlabel('Time')

plt.ylabel('Close Price')

plt.legend()

plt.show()



### LSTM Predictions vs True Values Plot

This plot compares the predictions made by a Long Short-Term Memory (LSTM) model against the true values of a time series over a given period.

* **True Values (blue line):** Represents the actual observed values of the time series.
* **LSTM Predictions (orange line):** Represents the values predicted by the LSTM model.

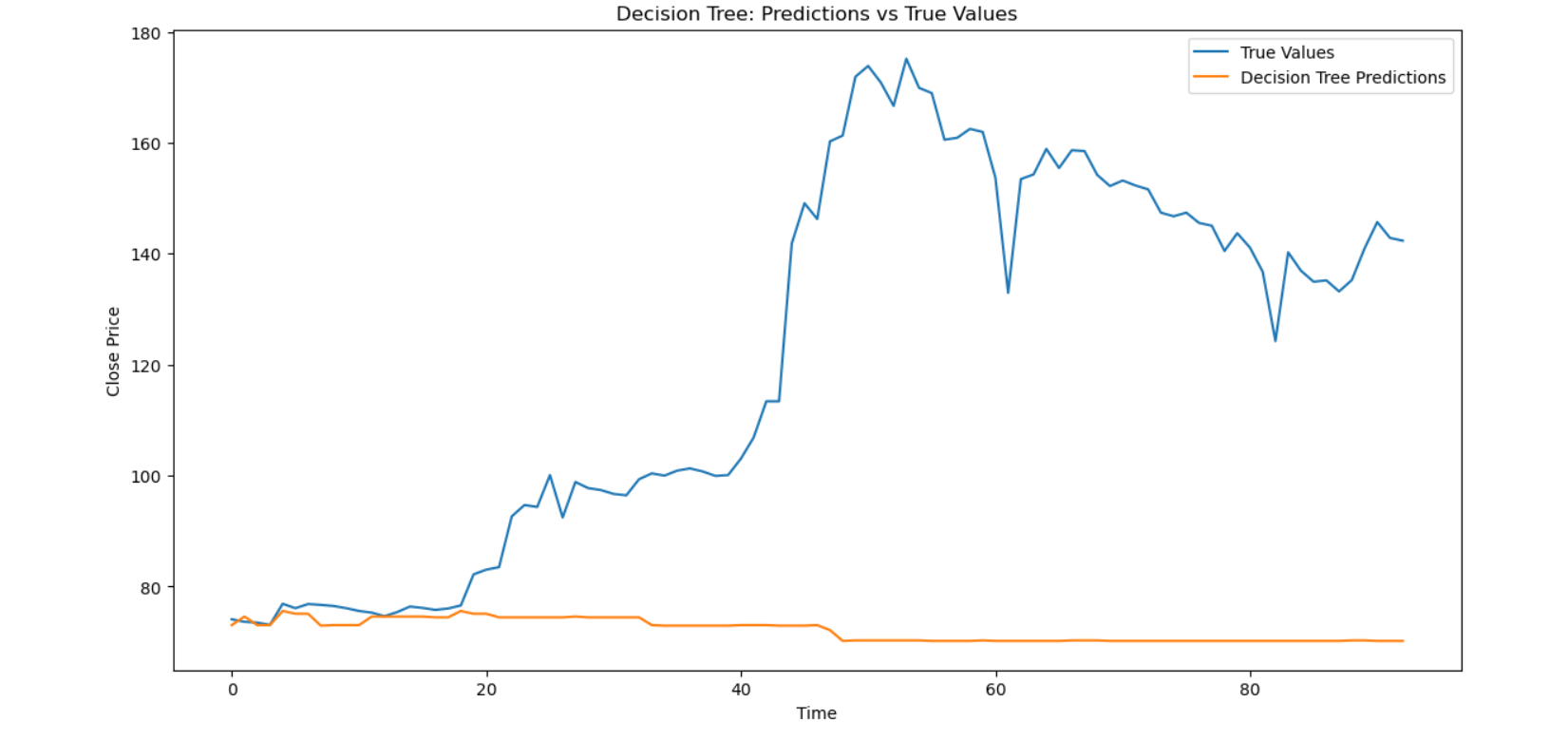
### Interpretation

1. **Initial Fit (Time 0 to 20):**
   * The LSTM predictions closely follow the true values, indicating a good fit in this early period.
2. **Middle Period (Time 20 to 50):**
   * The LSTM predictions still track the overall trend of the true values but begin to diverge slightly as the true values increase more rapidly.
   * The model captures the overall upward trend but underestimates the magnitude of the peaks.
3. **Later Period (Time 50 to 80):**
   * The divergence between the true values and LSTM predictions becomes more pronounced.
   * The true values exhibit more volatility, with significant fluctuations that the LSTM model does not capture accurately.
   * The LSTM predictions tend to smooth out the fluctuations and do not capture the sharp decreases observed in the true values.
4. **End Period (Time 80 onwards):**
   * The LSTM predictions continue to show a downward trend, while the true values have more pronounced fluctuations.

### Summary

* **Overall Trend:** The LSTM model captures the overall trend of the data reasonably well, especially in the initial and middle periods.
* **Prediction Accuracy:** The model's predictions are less accurate during periods of high volatility and sharp changes in the true values. The LSTM tends to smooth out the series and does not capture the peaks and troughs as well as the actual data.
* **Model Limitations:** This behavior suggests that while the LSTM model is good at capturing general trends, it may struggle with short-term volatility and abrupt changes. Further tuning or using additional features might improve its predictive performance.

These observations indicate that the LSTM model provides a good starting point for forecasting but may require further refinement to handle more volatile periods in the time series.



### Decision Tree: Predictions vs True Values Plot

This plot compares the predictions made by a Decision Tree model against the true values of a time series over a given period.

* **True Values (blue line):** Represents the actual observed values of the time series.
* **Decision Tree Predictions (orange line):** Represents the values predicted by the Decision Tree model.

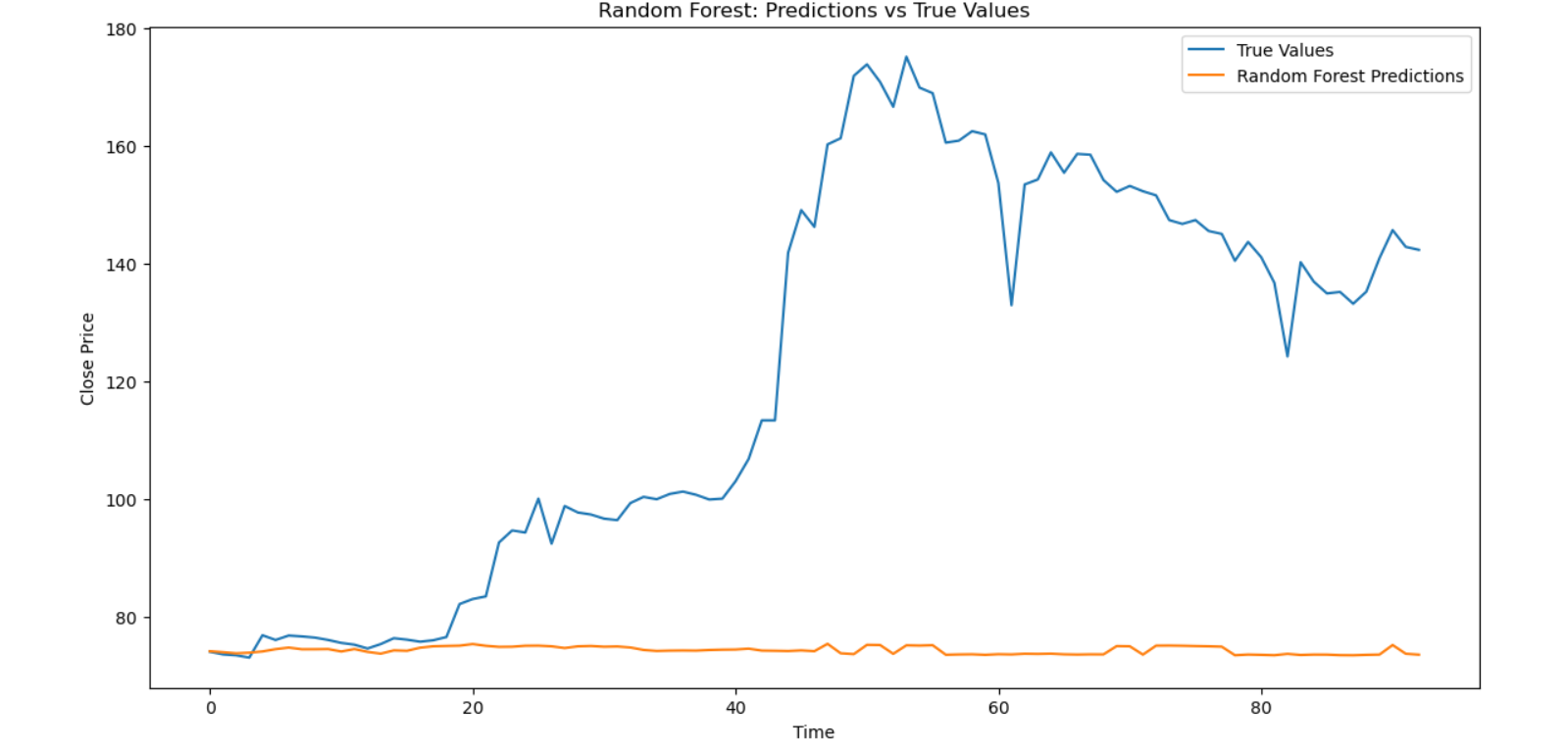
### Interpretation

1. **Initial Period (Time 0 to 20):**
   * The Decision Tree predictions somewhat follow the true values initially but fail to capture any significant upward trend.
   * The predicted values remain relatively flat compared to the true values that show some upward movement.
2. **Middle Period (Time 20 to 50):**
   * The true values exhibit a sharp increase, peaking around Time 40. However, the Decision Tree predictions do not reflect this increase and remain flat.
   * The model is not capturing the significant upward trend in the true values during this period.
3. **Later Period (Time 50 to 80):**
   * The true values display more fluctuations with a general downward trend.
   * The Decision Tree predictions remain flat and do not capture the volatility or the general downward trend of the true values.
4. **Overall Fit:**
   * The Decision Tree model consistently predicts values that are significantly lower and less variable compared to the true values.
   * The model fails to capture both the upward trends and the fluctuations observed in the true values.

### Summary

* **Overall Trend:** The Decision Tree model fails to capture the overall trend and variability in the true values.
* **Prediction Accuracy:** The model's predictions are flat and do not reflect the true values' patterns, especially during periods of sharp increase or volatility.
* **Model Limitations:** Decision Trees are known for their ability to handle non-linear relationships and interactions, but they can overfit the training data and fail to generalize well on unseen data, leading to poor performance on time series with trends and fluctuations.

The Decision Tree model appears to be unsuitable for this particular time series prediction task, likely due to its inability to capture the trend and volatility present in the data. Further model tuning or considering more complex models, such as ensemble methods (e.g., Random Forest, Gradient Boosting) or neural networks (e.g., LSTM), may yield better predictive performance.



### Random Forest: Predictions vs True Values Plot

This plot compares the predictions made by a Random Forest model against the true values of a time series over a given period.

* **True Values (blue line):** Represents the actual observed values of the time series.
* **Random Forest Predictions (orange line):** Represents the values predicted by the Random Forest model.

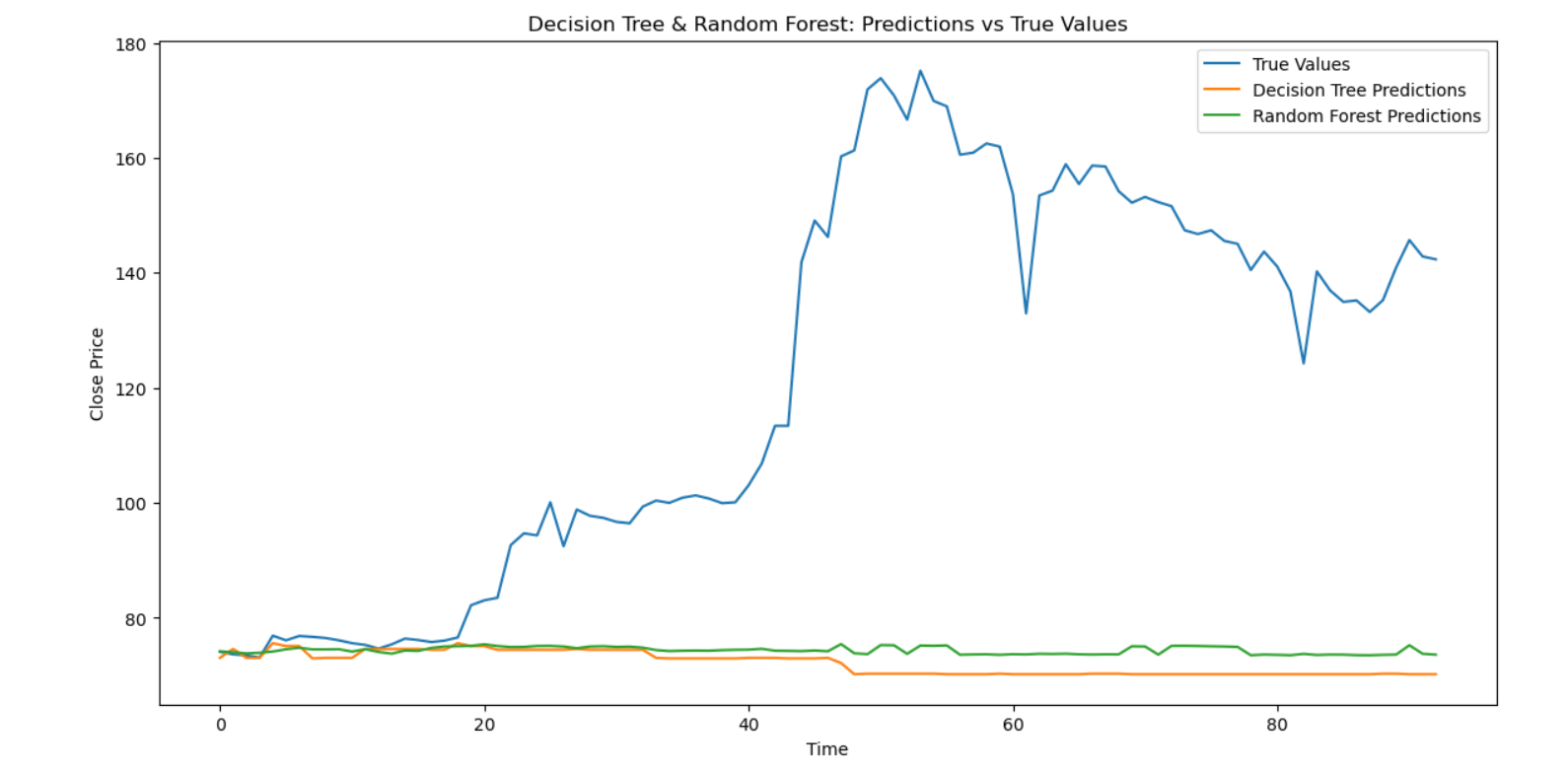
### Interpretation

1. **Initial Period (Time 0 to 20):**
   * The Random Forest predictions remain relatively flat compared to the true values, which show some initial variability and upward movement.
   * There is little alignment between the predicted and true values in this period.
2. **Middle Period (Time 20 to 50):**
   * The true values exhibit a sharp increase, peaking around Time 40. However, similar to the Decision Tree model, the Random Forest predictions remain flat and do not capture this increase.
   * The model is not capturing the significant upward trend in the true values during this period.
3. **Later Period (Time 50 to 80):**
   * The true values display more fluctuations with a general downward trend.
   * The Random Forest predictions continue to remain flat, not capturing the volatility or the general downward trend of the true values.
4. **Overall Fit:**
   * The Random Forest model, like the Decision Tree, consistently predicts values that are significantly lower and less variable compared to the true values.
   * The model fails to capture both the upward trends and the fluctuations observed in the true values.

### Summary

* **Overall Trend:** The Random Forest model fails to capture the overall trend and variability in the true values.
* **Prediction Accuracy:** The model's predictions are flat and do not reflect the true values' patterns, especially during periods of sharp increase or volatility.
* **Model Limitations:** Despite being an ensemble method that generally performs better than single Decision Trees, the Random Forest model appears to be unsuitable for this particular time series prediction task. It may be overfitting the training data or failing to capture the temporal dependencies in the data.

The Random Forest model's performance is not significantly better than the Decision Tree model. It suggests that more complex models or models specifically designed for time series data, such as LSTM networks, may be required to improve predictive performance.



The graph shows the comparison between true values and predictions made by a Decision Tree and a Random Forest model over time. Here are the key points to note:

1. **True Values (Blue Line)**: The true values exhibit a significant increase starting around time point 20, peaking near time point 40, and then showing a downward trend with some fluctuations thereafter.
2. **Decision Tree Predictions (Orange Line)**: The Decision Tree model's predictions are nearly flat and do not capture the upward trend observed in the true values. This indicates that the Decision Tree model is not effectively predicting the actual changes in the close price over time.
3. **Random Forest Predictions (Green Line)**: The Random Forest model's predictions are also relatively flat, similar to the Decision Tree predictions, and do not capture the upward trend in the true values.

### Interpretation

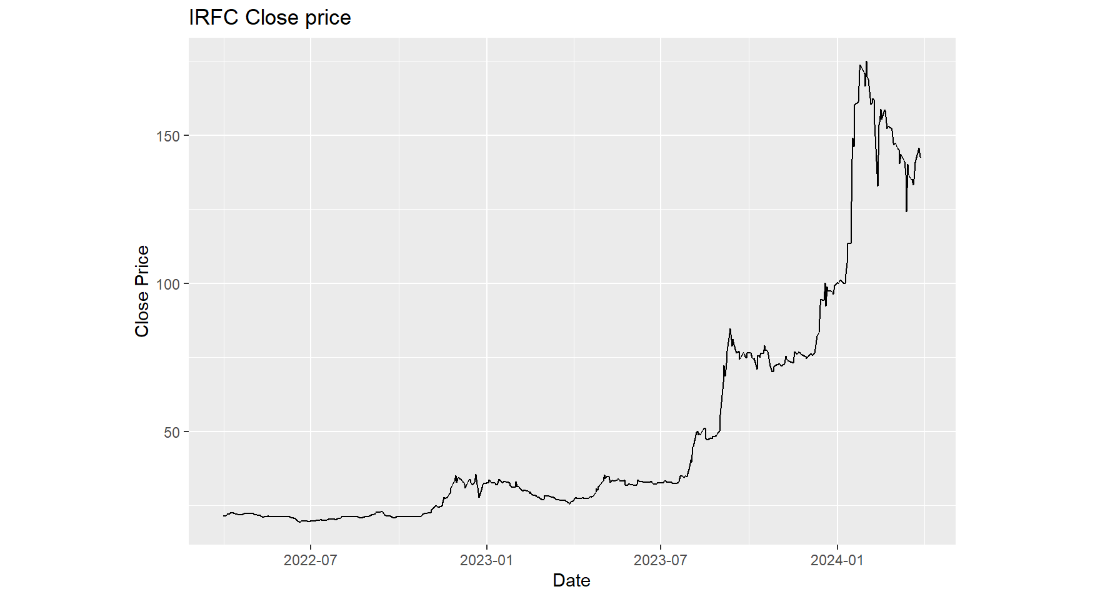
* **Model Performance**: Both the Decision Tree and Random Forest models are performing poorly in this scenario. They fail to capture the significant increase and subsequent fluctuations in the true values of the close price. This suggests that these models may not be suitable for predicting the given data or that they are not properly tuned.
* **Possible Causes for Poor Performance**:
  + **Model Complexity**: Decision Trees and Random Forests might not be complex enough to capture the intricate patterns in the data.
  + **Feature Engineering**: The features used for training the models might not be informative enough to predict the close price accurately.
  + **Overfitting/Underfitting**: The models might be overfitting to the training data or underfitting due to insufficient training.

### Recommendations

* **Feature Engineering**: Improve feature engineering to include more informative and relevant features.
* **Model Tuning**: Perform hyperparameter tuning for both models to improve their performance.
* **Advanced Models**: Consider using more advanced models such as Gradient Boosting Machines or Neural Networks, which might capture the patterns in the data more effectively.

INTERPRETATION AND RESULTS USING R:

Clean the data, check for outliers and missing values, interpolate the data if there are any missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data.



The graph displays the closing price of IRFC over a period extending from mid-2022 to early 2024. Here’s an interpretation of the data presented:

1. **Initial Period (Mid-2022 to Early 2023)**:
   * The closing price remains relatively stable, hovering around a low range.
   * There are minor fluctuations but no significant upward or downward trends.
2. **First Noticeable Increase (Early 2023)**:
   * Around early 2023, there is a noticeable increase in the closing price.
   * This increase is followed by a period of fluctuation but with a generally higher price level compared to the previous stable period.
3. **Major Surge (Mid-2023 to Early 2024)**:
   * Starting around mid-2023, there is a substantial surge in the closing price.
   * The price continues to climb steeply, reaching its peak around early 2024.
4. **Post-Peak Fluctuations (Early 2024)**:
   * After reaching the peak, the price experiences a sharp decline.
   * However, it stabilizes again at a level higher than the pre-surge prices but lower than the peak.

### Key Observations

* **Stability**: The initial phase is characterized by stability with minor fluctuations.
* **Growth Phase**: There are two distinct phases of growth: a moderate increase in early 2023 and a major surge starting mid-2023.
* **Volatility**: Post-peak, the price exhibits volatility but tends to stabilize at a higher level than during the initial period.

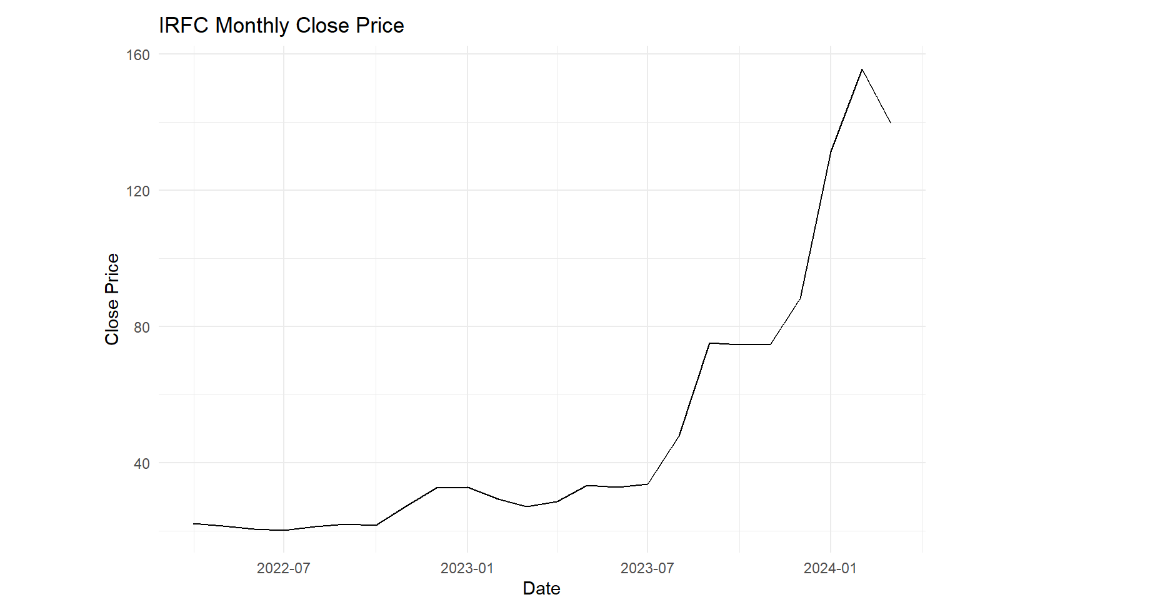
### Possible Causes for Trends

* **Initial Stability**: The initial stability could be due to consistent performance and market perception of IRFC.
* **Growth Phase**: The moderate and then major surge could be driven by positive market sentiment, improved financial performance, strategic initiatives, or broader market factors.
* **Post-Peak Fluctuations**: The volatility and subsequent stabilization could indicate market corrections, profit-taking by investors, or changes in market conditions.

### Recommendations for Investors

* **Initial Phase**: During stable periods, it might be advisable to monitor the market for any signs of upcoming trends or changes.
* **Growth Phase**: Investors might consider capitalizing on the upward trends while being cautious of potential peaks.
* **Post-Peak Volatility**: It’s essential to remain vigilant during volatile periods and consider market factors that might influence price stabilization or further declines.

Understanding the causes behind these trends requires a deeper analysis of IRFC’s financial reports, market conditions, and external economic factors during the specified period.



The graph shows the monthly closing price of IRFC from mid-2022 to early 2024. Here’s an interpretation of the data:

1. **Initial Period (Mid-2022 to Early 2023)**:
   * The closing price remains relatively low and stable, with minor fluctuations around the 30 to 40 range.
2. **Gradual Increase (Early 2023 to Mid-2023)**:
   * Around early 2023, the closing price begins to rise gradually.
   * This period shows a steady but slow upward trend in the closing price.
3. **Significant Surge (Mid-2023 to Early 2024)**:
   * Around mid-2023, there is a notable and sharp increase in the closing price.
   * The price rises significantly, reaching its peak around early 2024.
4. **Post-Peak Adjustment (Early 2024)**:
   * After reaching the peak, the closing price experiences a slight decline.
   * Despite the decline, the price remains much higher than the levels observed before mid-2023.

### Key Observations

* **Initial Stability**: The initial period is characterized by low and stable prices with minimal fluctuations.
* **Gradual Upward Trend**: The slow and steady increase in early 2023 suggests a gradual improvement in market conditions or investor sentiment.
* **Major Surge**: The sharp increase starting mid-2023 indicates a significant shift, possibly due to positive news, improved financial performance, or other market-driving factors.
* **Slight Decline**: The post-peak decline suggests a market correction or profit-taking by investors but still maintains a higher price level compared to the earlier periods.

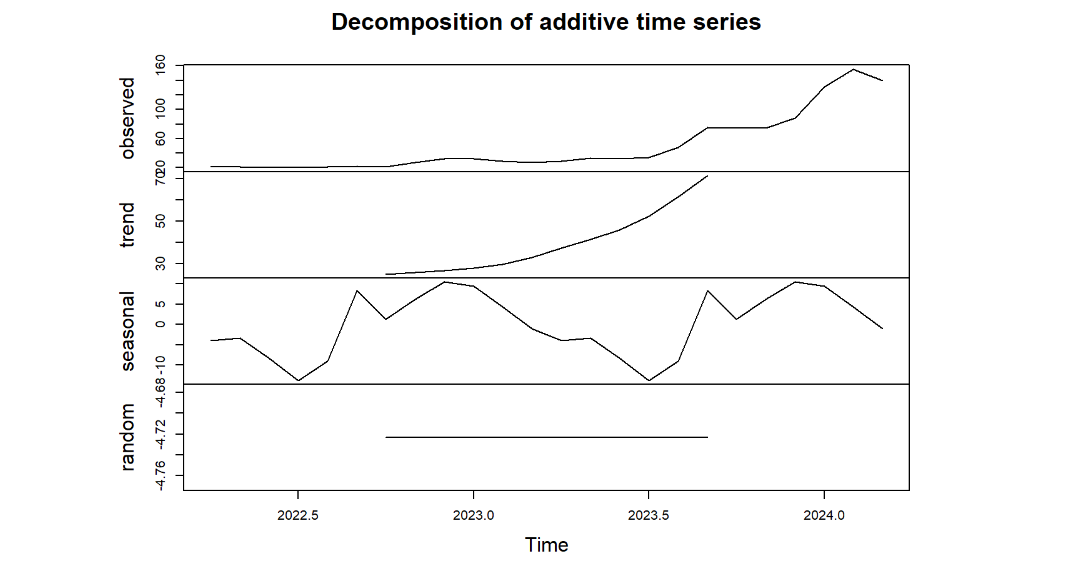
### Possible Causes for Trends

* **Initial Stability**: Consistent performance and market perception of IRFC, with no significant changes.
* **Gradual Increase**: Improved market sentiment, gradual accumulation of positive factors, or strategic initiatives starting to show impact.
* **Major Surge**: Major positive events, such as significant financial performance improvements, strategic announcements, or broader market trends positively impacting IRFC.
* **Post-Peak Adjustment**: Market correction, profit-taking by investors, or adjusting to new market conditions.

### Recommendations for Investors

* **During Stability**: Monitor the market for signs of upcoming changes or trends.
* **During Gradual Increase**: Consider gradual investment as the price begins to show a steady upward trend.
* **During Major Surge**: Capitalize on the sharp upward trend but remain cautious of potential peaks.
* **During Post-Peak Adjustment**: Be vigilant of further market corrections or stabilization, and consider the long-term potential of IRFC based on its new price level.

Understanding the underlying causes behind these trends requires a deeper analysis of IRFC’s financial performance, strategic initiatives, market conditions, and external economic factors during the specified period.



The image depicts the decomposition of an additive time series into its constituent components: observed, trend, seasonal, and random (residual) components.

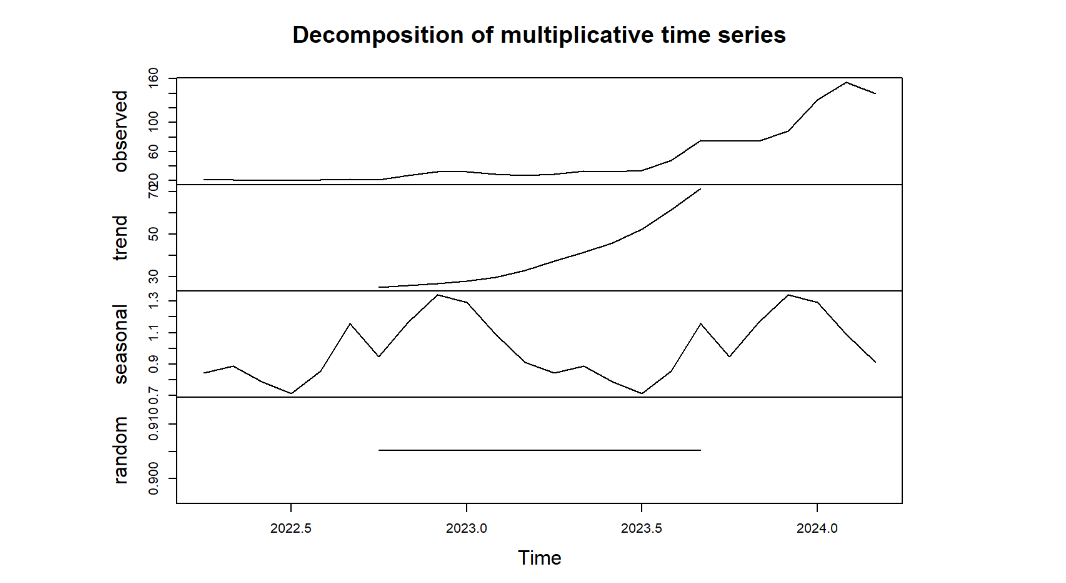
### Interpretation:

1. **Observed Component**:
   * The top plot represents the original observed time series data.
   * It shows the overall pattern of the data, including trends and fluctuations.
   * There is a clear upward trend observed from around mid-2023 onwards, with fluctuations throughout the entire period.
2. **Trend Component**:
   * The second plot represents the trend component of the time series.
   * It captures the long-term progression in the data, ignoring short-term fluctuations.
   * The trend starts flat and begins to increase steadily from around mid-2023, reflecting the overall upward movement in the observed data.
3. **Seasonal Component**:
   * The third plot shows the seasonal component, which represents the repeating short-term cycle within the data.
   * The seasonality appears to have periodic fluctuations with peaks and troughs occurring at regular intervals.
   * This component captures the cyclical patterns that repeat within a specific period (e.g., weekly, monthly).
4. **Random (Residual) Component**:
   * The bottom plot shows the random or residual component, representing the irregular, random fluctuations that are not explained by the trend or seasonal components.
   * This component should ideally have no discernible pattern and be centered around zero, indicating that the model has captured most of the systematic variation in the data.
   * In this plot, the residuals seem to be relatively stable around zero, with no significant patterns.

### Summary:

* **Observed**: The original data shows a significant upward trend starting from mid-2023, with observable fluctuations throughout the period.
* **Trend**: The trend component confirms the upward movement, particularly evident from mid-2023.
* **Seasonal**: The seasonal component reveals regular cyclical patterns, indicating periodic fluctuations in the data.
* **Random**: The random component shows the residual variations that are not captured by the trend or seasonal components, appearing relatively stable around zero.

This decomposition helps in understanding the underlying structure of the time series data, making it easier to analyze and forecast future values by addressing each component separately.



The image depicts the decomposition of a multiplicative time series into its constituent components: observed, trend, seasonal, and random (residual) components.

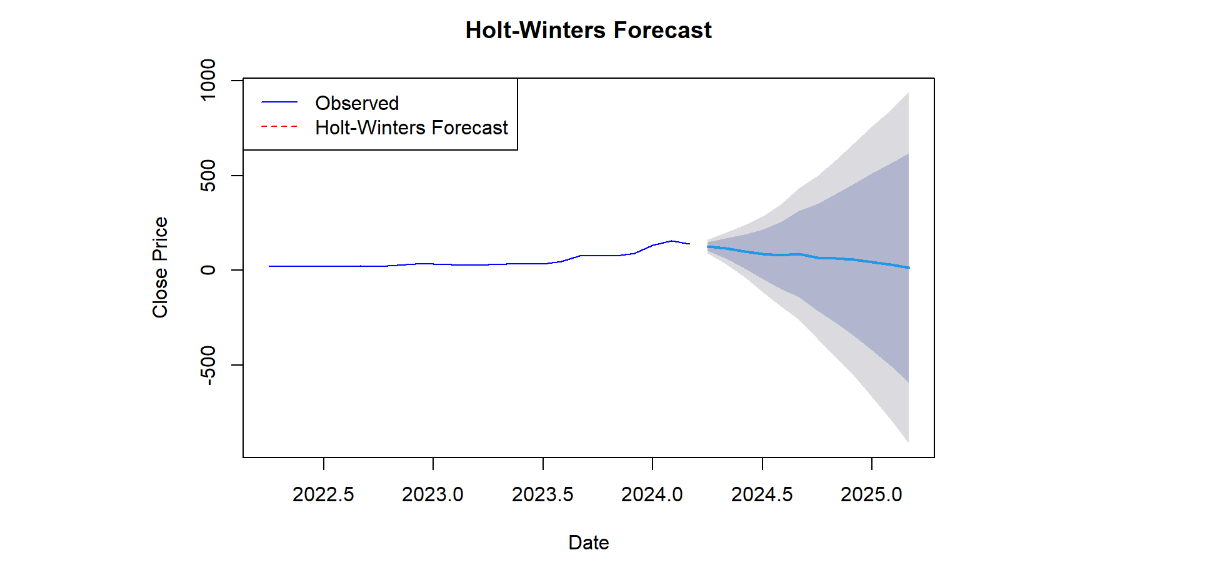
### Interpretation:

1. **Observed Component**:
   * The top plot represents the original observed time series data.
   * It shows the overall pattern of the data, including trends and fluctuations.
   * Similar to the additive decomposition, there is an initial period of relative stability followed by a significant upward trend starting around mid-2023, with observable fluctuations.
2. **Trend Component**:
   * The second plot represents the trend component of the time series.
   * It captures the long-term progression in the data, ignoring short-term fluctuations.
   * The trend shows a clear upward movement starting from mid-2023, reflecting the overall increase in the observed data.
3. **Seasonal Component**:
   * The third plot shows the seasonal component, representing the repeating short-term cycle within the data.
   * The seasonal component in a multiplicative decomposition is shown as a ratio around 1. Values above 1 indicate that the seasonality is boosting the data, while values below 1 indicate that it is reducing the data.
   * The seasonality shows regular fluctuations, indicating periodic patterns within the data.
4. **Random (Residual) Component**:
   * The bottom plot shows the random or residual component, representing the irregular, random fluctuations not explained by the trend or seasonal components.
   * This component should ideally have no discernible pattern and be centered around 1 in a multiplicative model, indicating that the model has captured most of the systematic variation in the data.
   * The residuals appear relatively stable, with slight deviations around the mean.

### Summary:

* **Observed**: The original data shows a significant upward trend starting from mid-2023, with observable fluctuations throughout the period.
* **Trend**: The trend component confirms the upward movement, particularly evident from mid-2023.
* **Seasonal**: The seasonal component reveals regular cyclical patterns, indicating periodic fluctuations that either amplify or diminish the observed values.
* **Random**: The random component shows the residual variations that are not captured by the trend or seasonal components, appearing relatively stable around 1.

This decomposition helps in understanding the underlying structure of the time series data, allowing for more accurate analysis and forecasting by addressing each component separately.



The image shows a time series plot of observed data along with a Holt-Winters forecast.

### Interpretation:

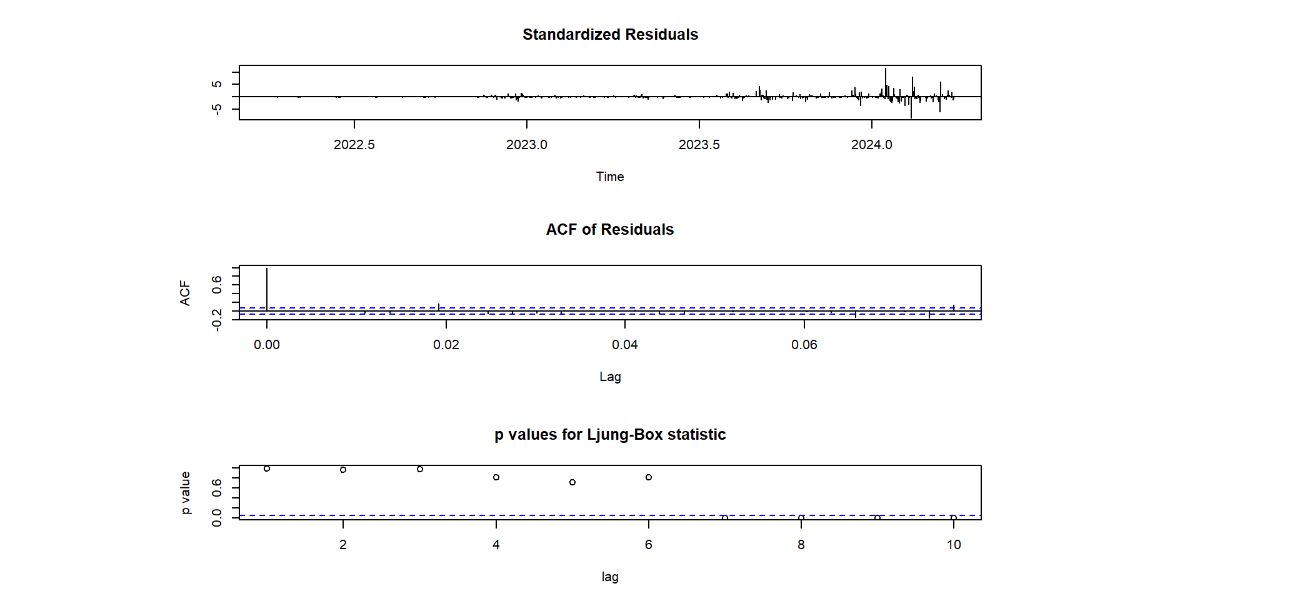
1. **Observed Data**:
   * The blue line represents the historical observed data of the close price from around early 2022 to early 2024.
   * The data remains relatively stable throughout this period, with minor fluctuations around a constant value.
2. **Holt-Winters Forecast**:
   * The red dashed line represents the Holt-Winters forecast for the close price beyond the historical data.
   * The forecast starts from early 2024 and extends to early 2025.
   * The Holt-Winters forecast predicts a continuation of the relatively stable trend observed in the historical data.
3. **Prediction Interval**:
   * The shaded area around the forecast represents the prediction interval, which gives a range within which the future values are expected to fall.
   * The light grey area represents a wider prediction interval, while the dark grey area represents a narrower prediction interval.
   * As the forecast extends further into the future, the prediction interval widens, indicating increasing uncertainty.

### Key Observations:

* The observed data shows a relatively stable trend with minor fluctuations.
* The Holt-Winters forecast suggests that this stability will continue, with the close price expected to remain around the same level.
* The prediction intervals widen over time, reflecting greater uncertainty in the forecast as it projects further into the future.

### Summary:

* The historical data shows minor fluctuations around a relatively constant value.
* The Holt-Winters forecast predicts that this stability will continue, with the close price remaining relatively unchanged.
* The widening prediction intervals indicate increasing uncertainty in the forecast over time, which is a common characteristic of time series forecasts.



The image consists of three diagnostic plots for a time series model: standardized residuals, autocorrelation function (ACF) of residuals, and p-values for the Ljung-Box statistic.

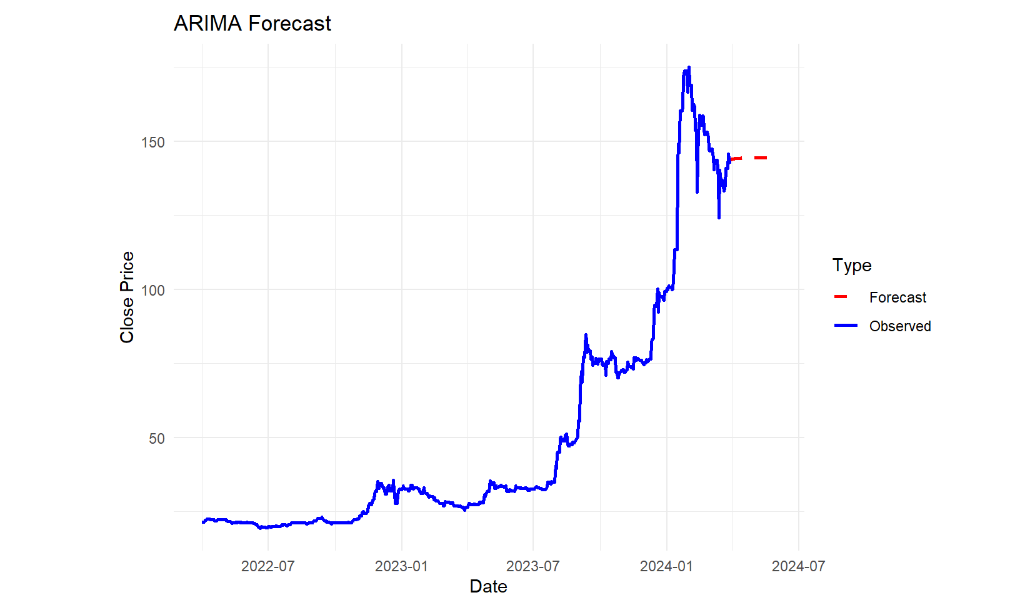
### Interpretation:

1. **Standardized Residuals**:
   * The top plot shows the standardized residuals over time.
   * The residuals should ideally be randomly scattered around zero with no clear patterns if the model fits the data well.
   * **Observation**:
     + The residuals are mostly centered around zero with no significant patterns, but there are some periods of increased volatility, particularly towards the end of the series.
2. **ACF of Residuals**:
   * The middle plot shows the autocorrelation function of the residuals.
   * The x-axis represents the lag (time steps), and the y-axis represents the autocorrelation.
   * The blue dashed lines represent the 95% confidence intervals.
   * **Observation**:
     + Most of the autocorrelation values fall within the confidence intervals, indicating that the residuals are mostly uncorrelated.
     + There is a significant spike at lag 0, which is expected, but subsequent lags do not show significant autocorrelation, suggesting that the residuals do not exhibit serial correlation.
3. **p-values for Ljung-Box Statistic**:
   * The bottom plot shows the p-values for the Ljung-Box Q test, which tests for the presence of autocorrelation in the residuals.
   * The x-axis represents the lag, and the y-axis represents the p-value.
   * The blue dashed line represents the significance level (commonly 0.05).
   * **Observation**:
     + Most of the p-values are above the significance level (0.05), indicating that there is no significant autocorrelation in the residuals at these lags.
     + This suggests that the model adequately captures the autocorrelation structure of the time series data.

### Summary:

* **Standardized Residuals**: The residuals are mostly centered around zero with no significant patterns, although there is some increased volatility towards the end of the series.
* **ACF of Residuals**: The residuals are mostly uncorrelated, with most autocorrelation values within the 95% confidence intervals.
* **p-values for Ljung-Box Statistic**: The p-values are mostly above the significance level, indicating no significant autocorrelation in the residuals at the tested lags.

Overall, these diagnostics suggest that the time series model fits the data reasonably well, with no major issues of autocorrelation in the residuals.



The image shows a time series plot of observed data along with an ARIMA (AutoRegressive Integrated Moving Average) forecast.

### Interpretation:

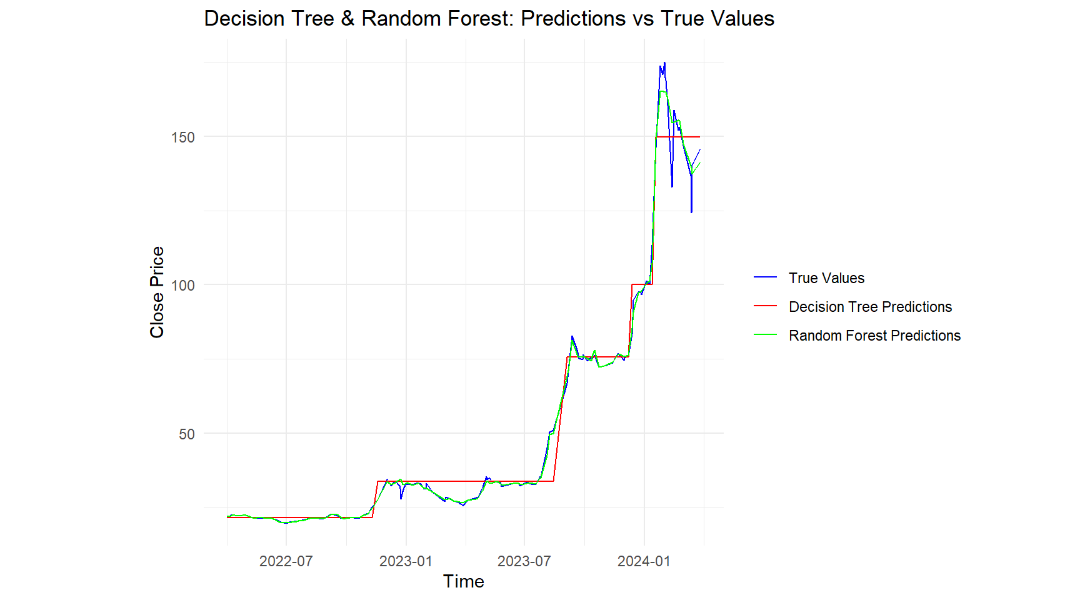
1. **Observed Data**:
   * The blue line represents the historical observed data of the close price from around mid-2022 to early 2024.
   * The observed data shows a period of relative stability, followed by a significant upward trend starting around the beginning of 2023, with noticeable fluctuations and volatility.
2. **ARIMA Forecast**:
   * The red dashed line represents the ARIMA forecast for the close price beyond the historical data.
   * The forecast starts from early 2024 and extends slightly into the future, providing a short-term prediction.
   * The ARIMA model predicts a relatively stable close price in the near future, with slight fluctuations around the current observed values.

### Key Observations:

* The observed data exhibits a sharp increase in the close price starting from early 2023, reaching a peak, followed by a slight decline and stabilization towards the end of the observed period.
* The ARIMA forecast suggests that the close price will remain relatively stable in the short term, without significant increases or decreases.
* The forecast reflects the recent stabilization observed in the data, indicating that the model anticipates this trend to continue.

### Summary:

* The historical data shows a period of stability, followed by a sharp increase in the close price with significant volatility.
* The ARIMA model forecasts that the close price will stabilize around the current level in the near future, with only minor fluctuations.
* This analysis provides a short-term outlook based on the recent trends in the observed data, which can be useful for immediate planning and decision-making.



The image shows a time series plot comparing true values of close prices with predictions made by Decision Tree and Random Forest models.

### Interpretation:

1. **True Values (Blue Line)**:
   * The blue line represents the actual observed close prices from around mid-2022 to early 2024.
   * The data shows a period of stability followed by a significant upward trend starting around the beginning of 2023, with noticeable volatility and fluctuations.
2. **Decision Tree Predictions (Red Line)**:
   * The red line represents the predictions made by the Decision Tree model.
   * Decision Tree predictions are piecewise constant, capturing some of the major changes in the trend but missing finer fluctuations.
   * The model seems to capture the general upward trend but with some noticeable deviations from the true values, especially during periods of rapid change.
3. **Random Forest Predictions (Green Line)**:
   * The green line represents the predictions made by the Random Forest model.
   * The Random Forest model provides a smoother prediction, closely following the true values and capturing both the major trends and finer fluctuations better than the Decision Tree model.
   * The predictions align more closely with the observed data, indicating better performance in capturing the underlying patterns.

### Key Observations:

* **True Values**: The actual observed data shows a significant increase in the close price starting from early 2023, with notable volatility.
* **Decision Tree Predictions**: The Decision Tree model captures the general trend but struggles with the finer details and rapid changes in the data, leading to some deviations.
* **Random Forest Predictions**: The Random Forest model provides a closer fit to the true values, capturing both the overall trend and the finer fluctuations more accurately.

### Summary:

* The observed data shows a period of stability followed by a significant upward trend with volatility.
* The Decision Tree model captures the general trend but misses finer fluctuations and rapid changes.
* The Random Forest model performs better, providing smoother and more accurate predictions that align closely with the true values.
* Overall, the Random Forest model appears to be more effective in modeling the close prices in this time series, offering better predictive performance compared to the Decision Tree model.

Top of Form

Bottom of Form