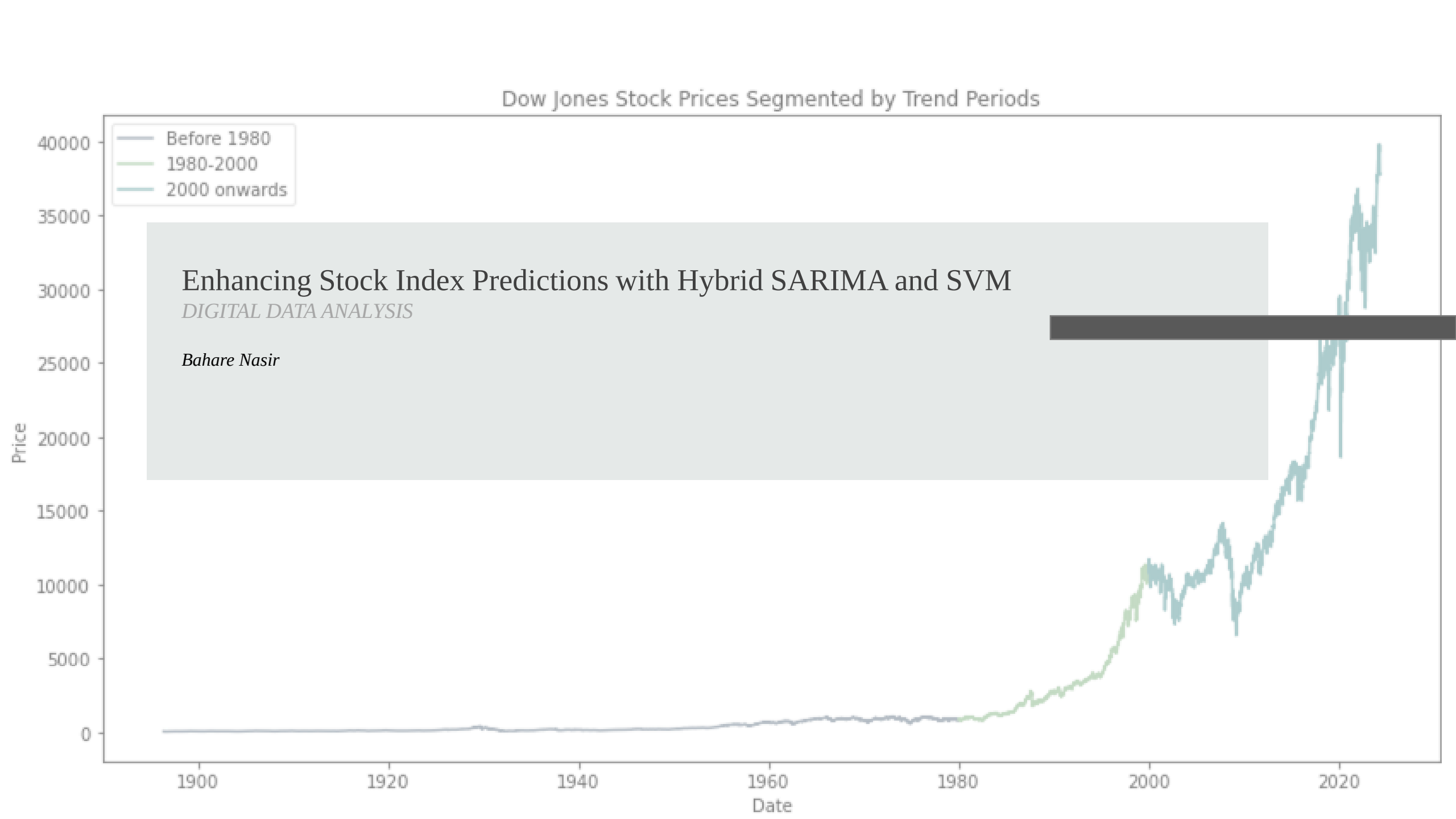


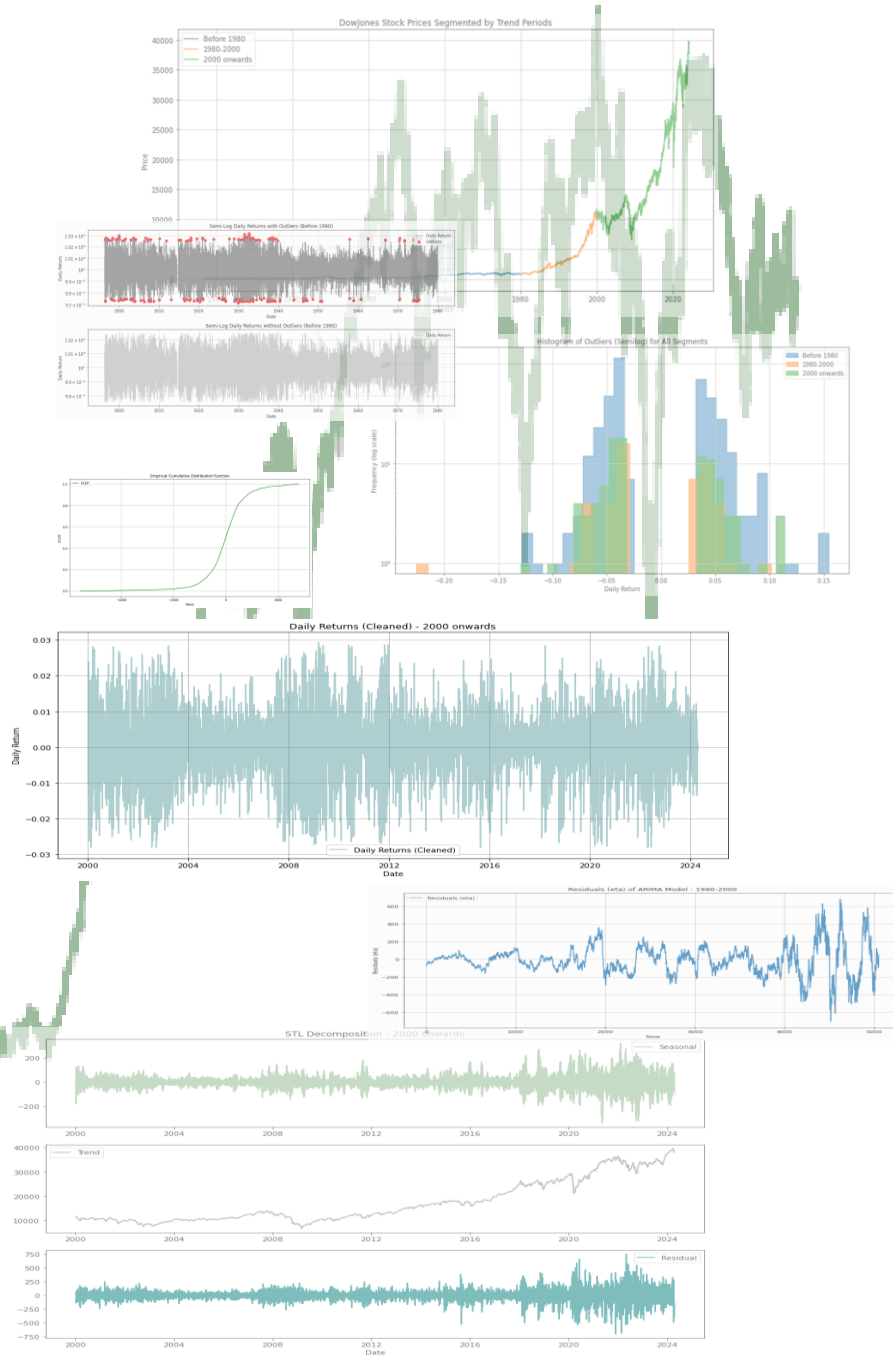
Enhancing Stock Index Predictions with Hybrid SARIMA and SVM

DIGITAL DATA ANALYSIS

Bahare Nasir

Context:

- Introduction
- Suggested research methodology
- Data Preparation
- Anomaly Detection Using Matrix Profiles
- Statistical Analysis
- STL Decomposition
- The Dickey-Fuller Test
- Kolmogorov-Smirnov Test
- Introducing Models
- Forecasting
- Conclusion
- Future Work Suggestions
- References



Introduction:

The Dynamics of the Stock Market:

Stock Market:

- Dynamics platform for issuing, buying, and selling securities
- Allows companies to raise capital, investors to earn returns

Influences:

- Economic indicators
- Corporate performance
- Geopolitical events

Introducing the importance of the Data:

Dow Jones Industrial Average:

- Closely watched stock market index

Insights from Historical Data:

- Long-term economic trends
- Market cycles
- Impacts of major events

Benefits:

- Comprehensive market behavior analysis
- Enhanced understanding of stock market dynamics

Understanding Time Series Analysis:

Time series analysis is crucial for stock market forecasting, identifying patterns and predicting future price movements. Integrating traditional ARIMA models with advanced SVM methods enhances forecast accuracy and decision-making in complex financial markets.

Key Techniques:

- ARIMA/SARIMA Models:
  - capture linear patterns and seasonal trends variations
  - Useful for forecasting
- SVM (Support Vector Machine):
  - captures complex, non-linear relationships in data
  - Improves accuracy by addressing dynamics missed by traditional models.

Linking Time Series Analysis and Stock Market Dynamics:

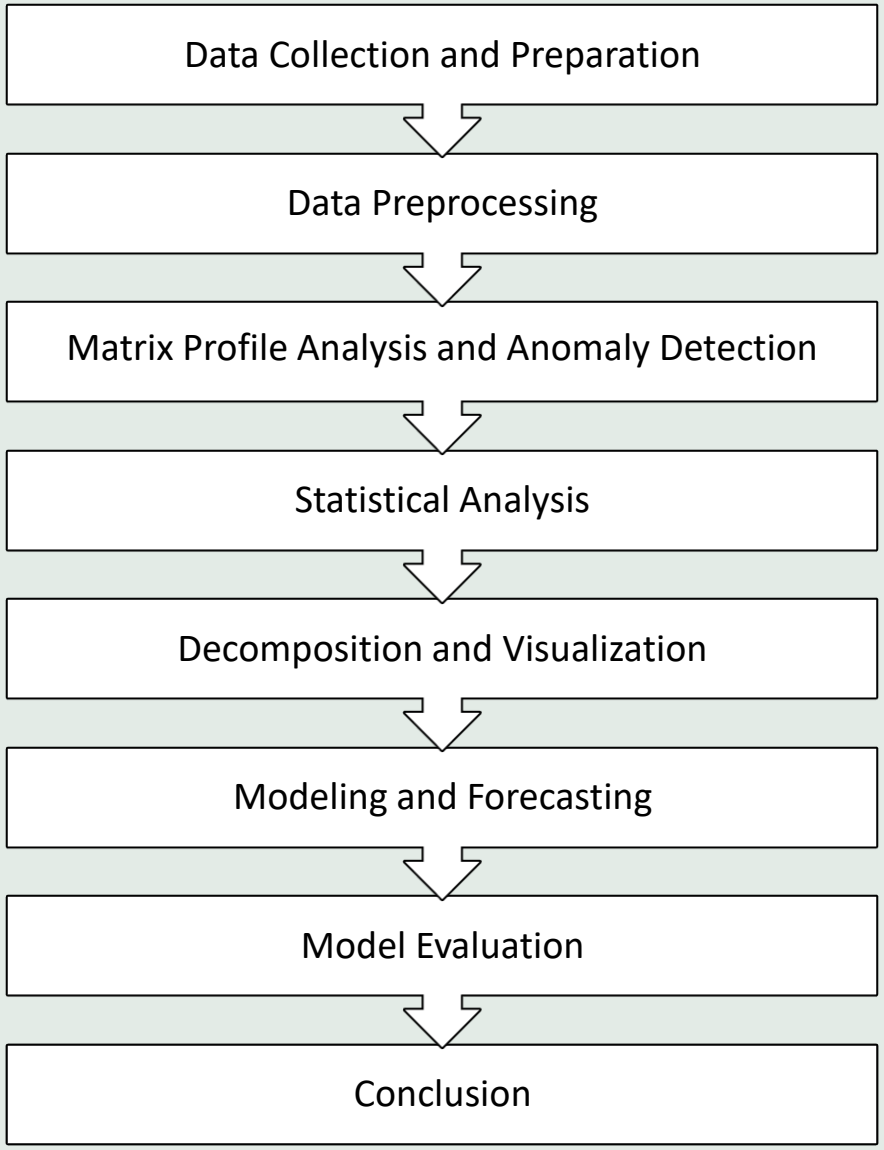
- Pattern Recognition and Forecasting:

Time series analysis helps in identifying trends, seasonality, and anomalies in stock market data, which are crucial for forecasting future price movements.

- Hybrid Models for Better Accuracy:

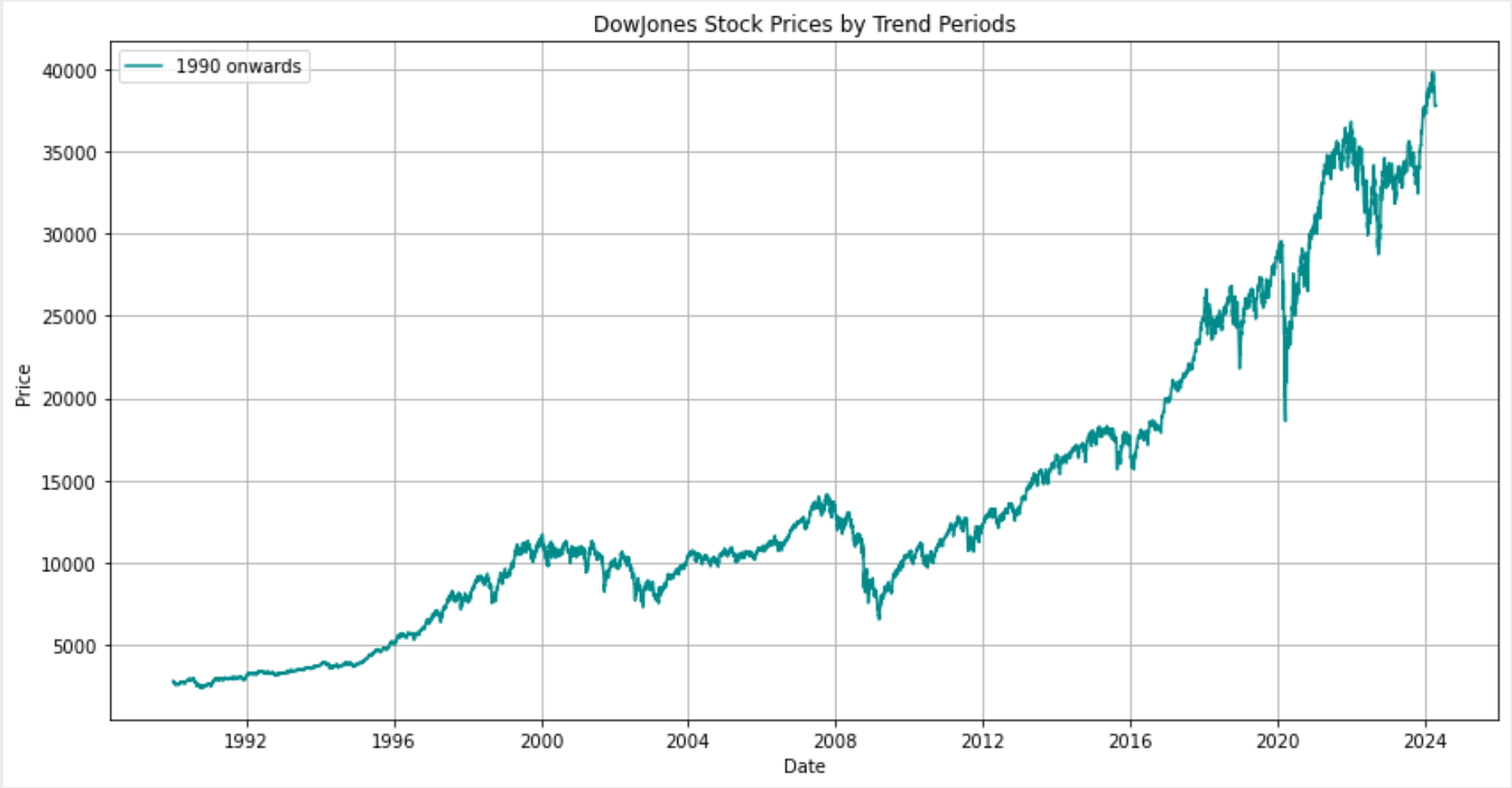
Combining traditional models like ARIMA/SARIMA with advanced machine learning methods such as SVM enhances the ability to predict both linear and non-linear market behaviors, leading to more robust and reliable investment decisions.

Suggested research methodology:



Data Preparation:

Analyze market behaviors and trends over time:



Data Composition:

Our dataset consists of historical stock prices, specifically the closing prices of the Dow Jones index, which reflect the final price at which a stock is traded on a particular trading day. However, analyzing raw closing prices directly can present challenges due to the presence of trends, seasonal effects, and varying volatility over time.

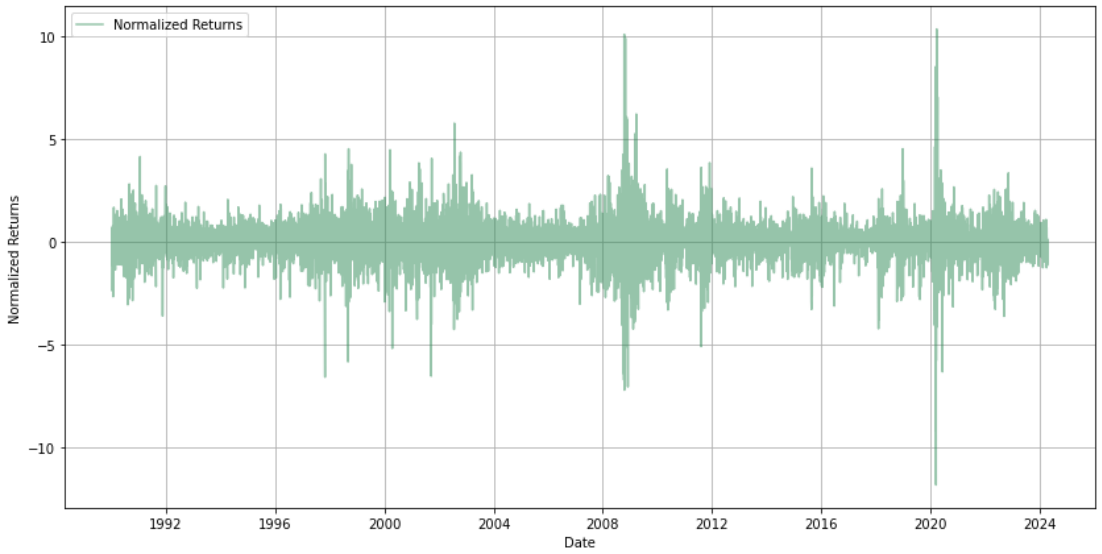
Why Not Use Close Prices Directly:

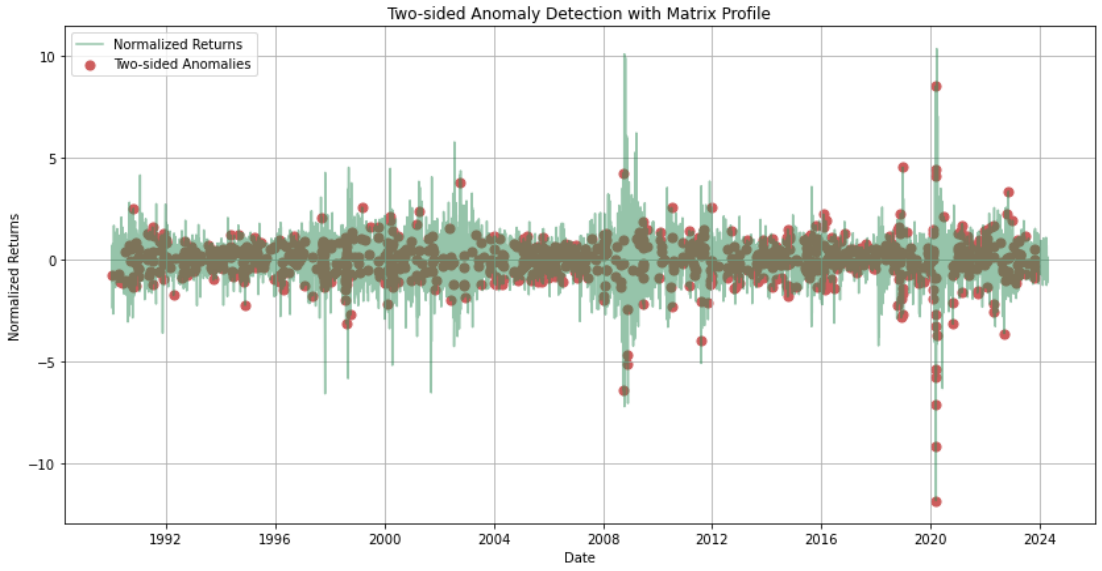
- Non-Stationarity: Closing prices often show trends and changing scales, making them difficult to model and predict accurately.
- Volatility Misrepresentation: High prices can appear more volatile than lower ones, even with identical percentage changes.

Why We Use Returns:

- Achieving Stationarity: Returns, calculated as percentage changes, help remove trends and stabilize variance, making data more suitable for models.
- Relative Changes: Returns focus on percentage changes, making it easier to compare across different stocks and periods.
- Similar to Differencing: Returns act like differencing, capturing daily price movements without long-term biases.

Anomaly Detection in Time Series Data Using Matrix Profiles:





Data Collection and Preparation:

- Filter data from 1990 onwards and calculate daily returns to assess stock price variability.

Normalization:

- Use Z-score normalization (StandardScaler). to normalize returns, ensuring consistent scale for comparison.

Matrix Profile Calculation and Anomaly Detection:

- Calculate matrix profiles to identify discords, which are the most anomalous subsequences.

Two-Sided Anomaly Detection:

- Set upper and lower thresholds (mean ± 1.5 * std) to detect significant deviations in both directions.

Statistical Analysis:

- Analyze mean, standard deviation, skewness, and kurtosis to understand data distribution with anomalies.

Outlier Replacement:

- Replace anomalies with the median value to mitigate the impact of outliers.

Visualization:

- Plot normalized returns and anomalies to validate detection effectiveness and threshold choices

Why Two-Sided Anomaly Detection?

- Financial Context:
  Detects both positive (sharp rises) and negative (sudden falls) anomalies in stock prices.
- Balanced Sensitivity:
  Captures deviations on both ends, ensuring sensitivity to all types of anomalies.
- Versatility:
  Flexible and adaptable to contexts requiring detection of both upper and lower anomalies.

Why Not Other Methods?

- Specificity:
  Matrix profiles provide detailed similarity measures of subsequences, ideal for time series analysis.
- Efficiency:
  Computationally efficient, allowing precise detection of both local and global anomalies.
- Comparative Simplicity:
  Offers a direct, interpretable approach without extensive tuning or training, unlike machine learning methods.

Statistical Analysis:

Purpose:

To assess the impact of replacing outliers with the median on the distribution of returns.

Before Replacing Outliers (with Anomalies):

- Mean:0.0361
- Standard Deviation: 0.010
- Skewness: -0.308 (slightly skewed left)
- Kurtosis: 12.303 (high kurtosis, indicating heavy tails)

After Replacing Outliers with Median:

- Mean: 0.05 (slightly increased)
- Standard Deviation: 0.008 (reduced variability)
- Skewness: 0.208 (shifted to slightly skewed right)
- Kurtosis: 10.178 (still high but reduced)

Benefits of Using Median for Outlier Replacement:

- **Robustness:**

  less affected by extreme values, providing a stable central value.

- **Preserves Data:**

  Keeps all data points, unlike removal, avoiding loss of information.

- **Minimizes Distortion:**

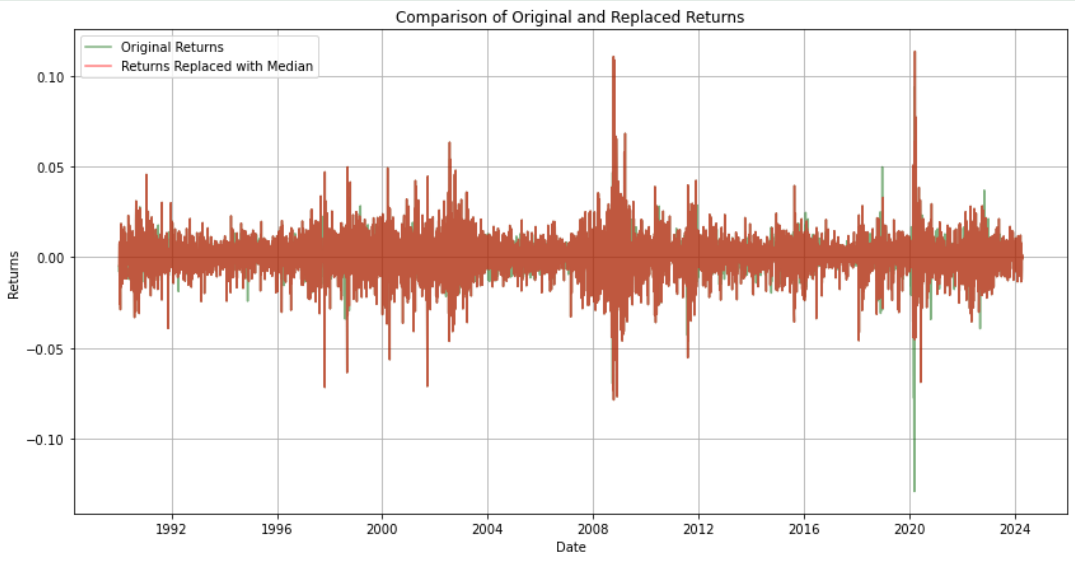
  More accurate than the mean, which can still be skewed by outliers.

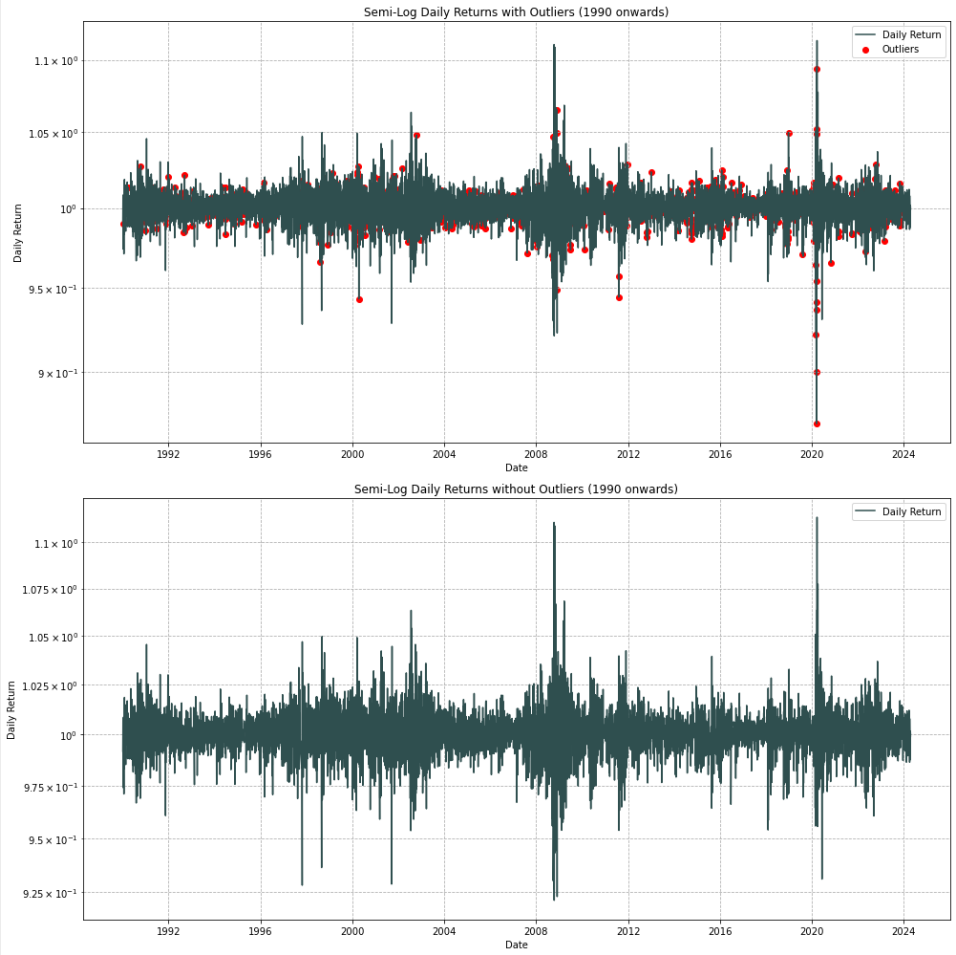
- **Balances Distribution:**

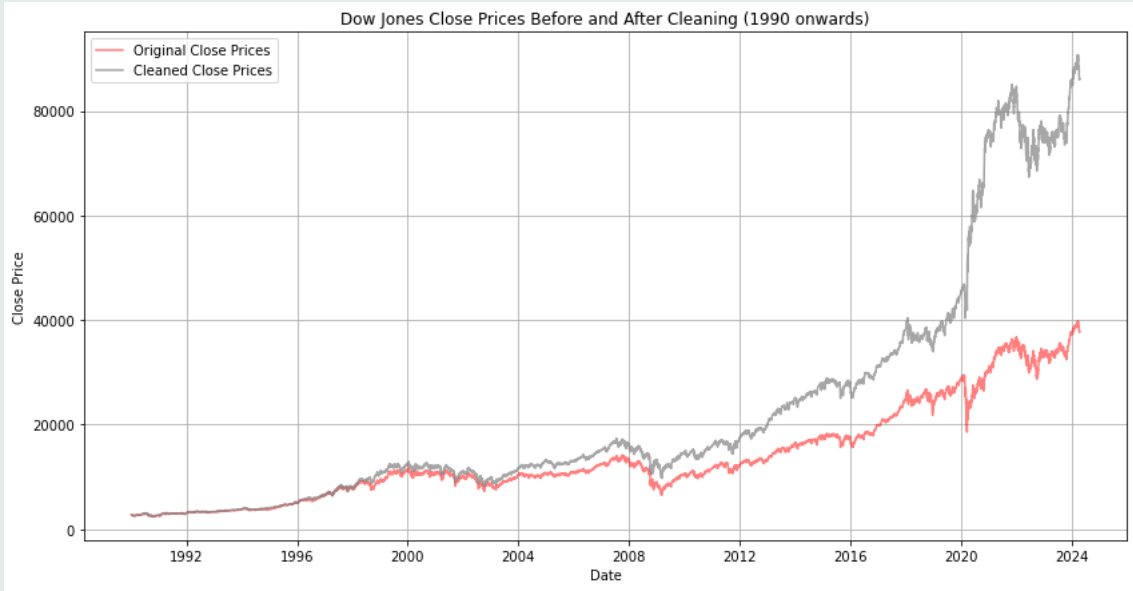
  Reduces the impact of extremes, leading to a more balanced dataset.

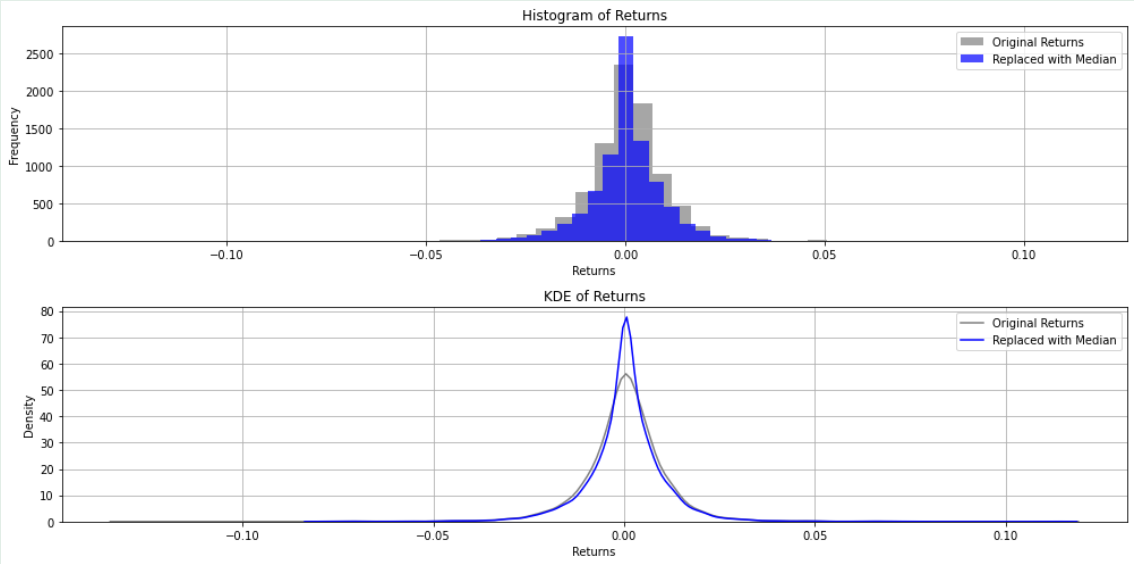
- **Maintains Order:**

  Preserves the sequential integrity of time series data.




Semi-Log Daily Returns with Outliers (1990 onwards)
Daily Return
Outliers
Daily Return
Date
Semi-Log Daily Returns without Outliers (1990 onwards)
Daily Return
Daily Return
Date


Dow Jones Close Prices Before and After Cleaning (1990 onwards)
Original Close Prices
Cleaned Close Prices
Close Price
Date

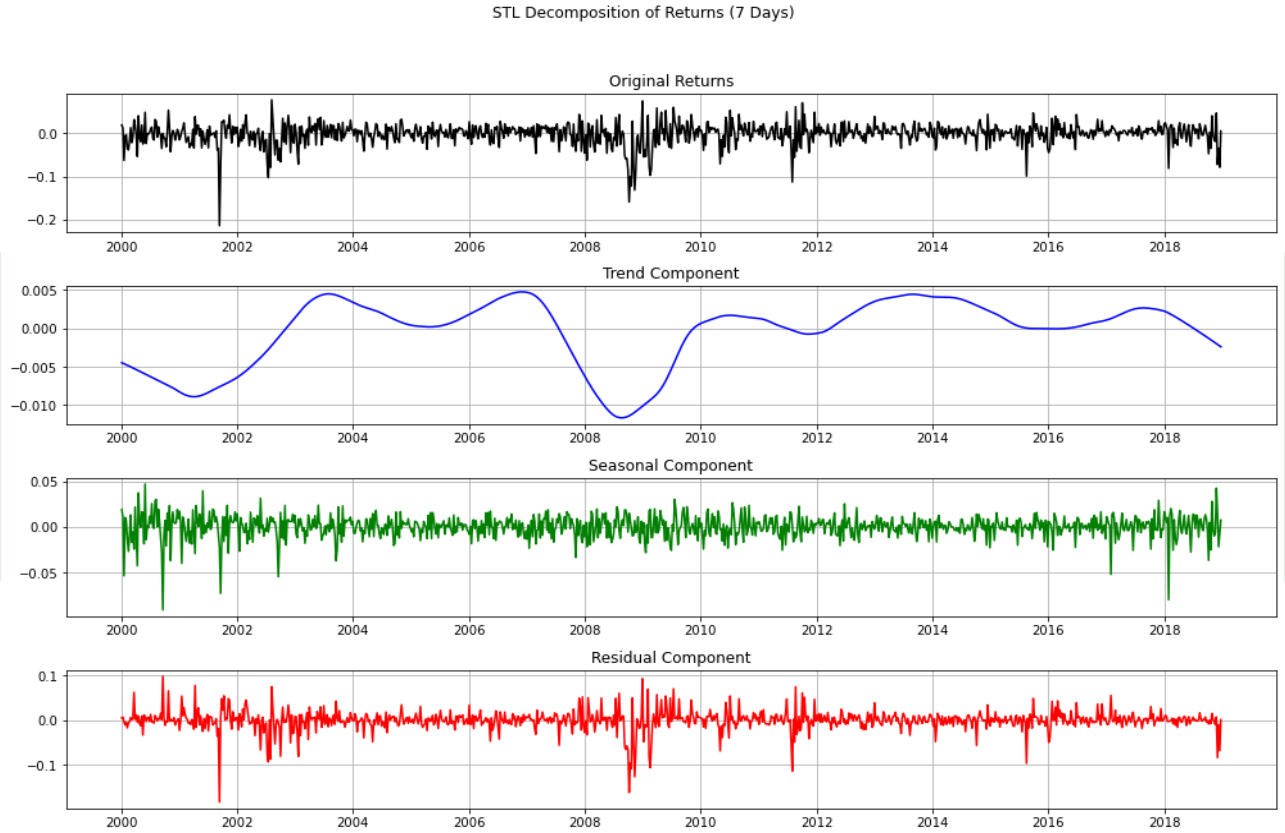

Histogram of Returns
Original Returns
Replaced with Median
Frequency
Returns
KDE of Returns
Original Returns
Replaced with Median
Density
Returns

STL Decomposition of Returns (7 Days):

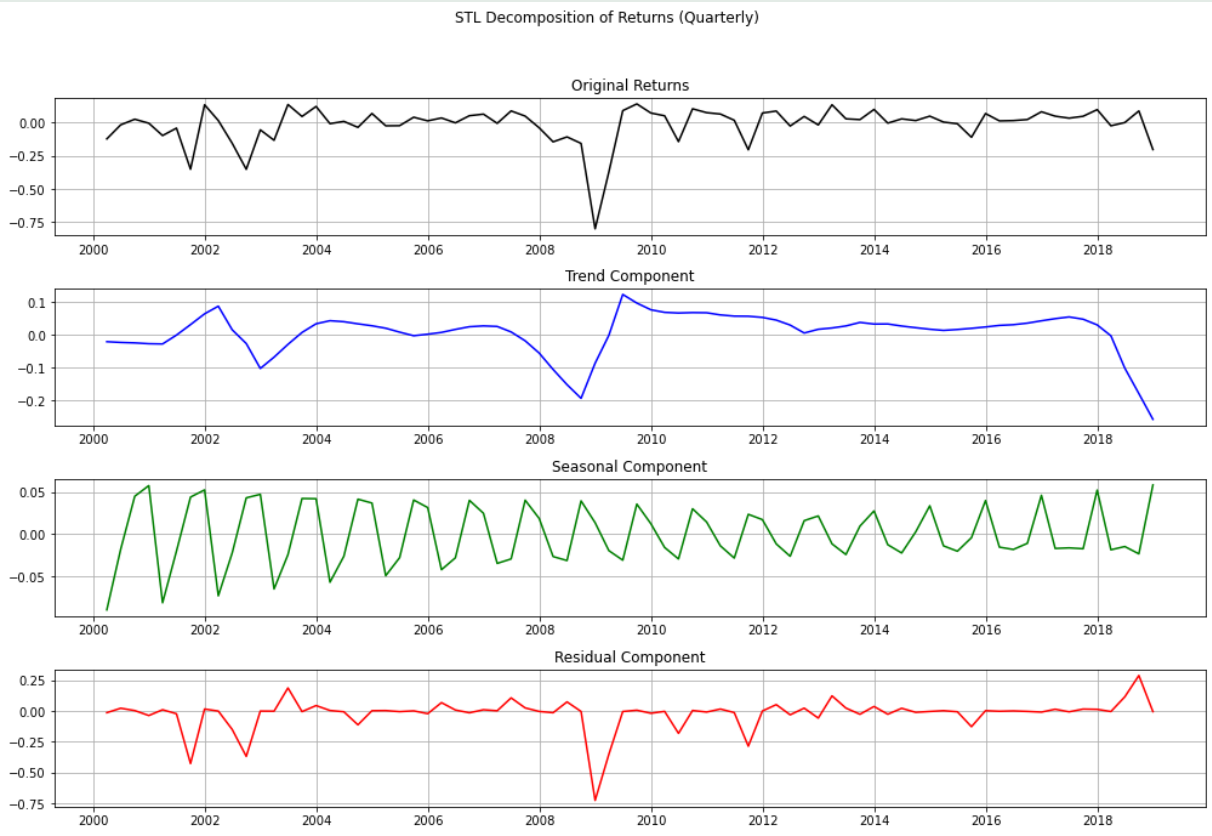
Period: 7 days

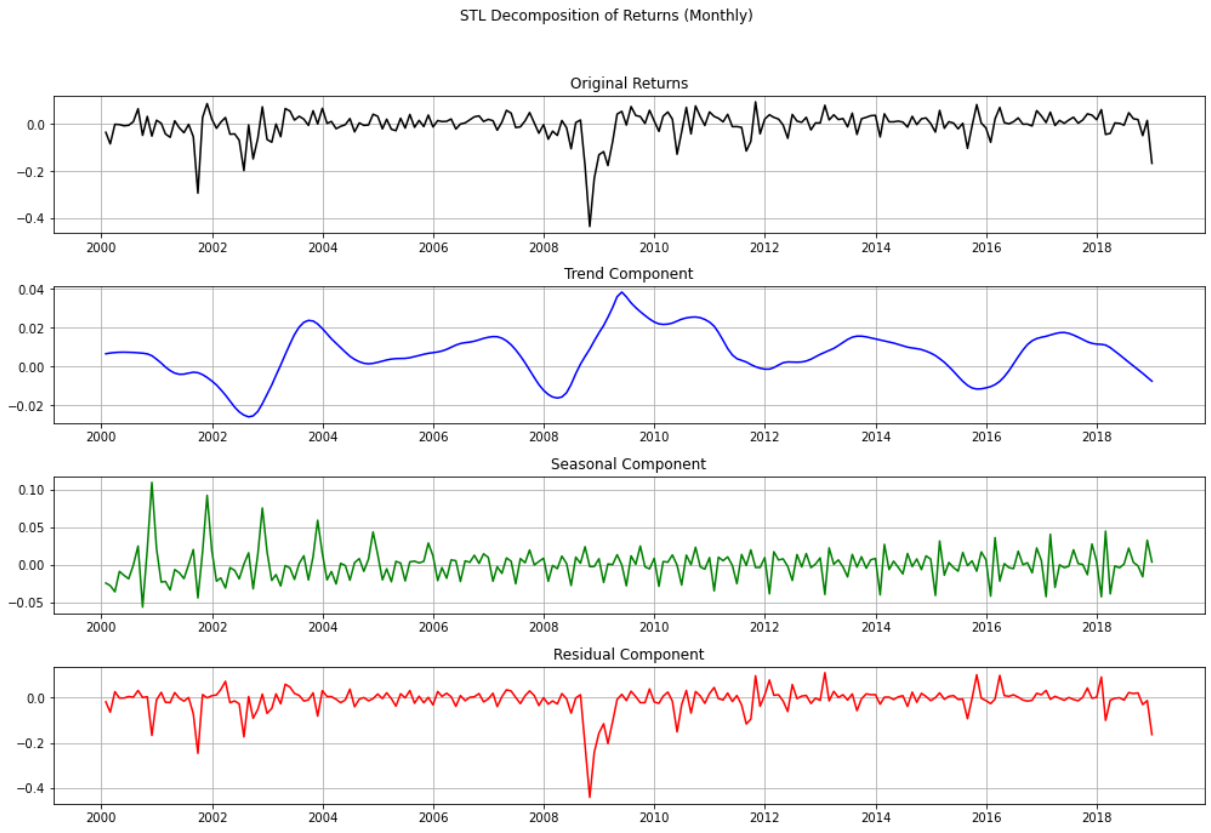
Original Returns	Trend	Seasonal Component	Residuals
• Shows daily market fluctuations	• Highlights the overall long-term direction with cyclical movements.	• Captures consistent weekly patterns in returns.	• Reflect random noise and unexpected market events.

This decomposition helps identify long-term trends, recurring weekly cycles, and irregular market behaviors, aiding in better forecasting and analysis of stock returns.



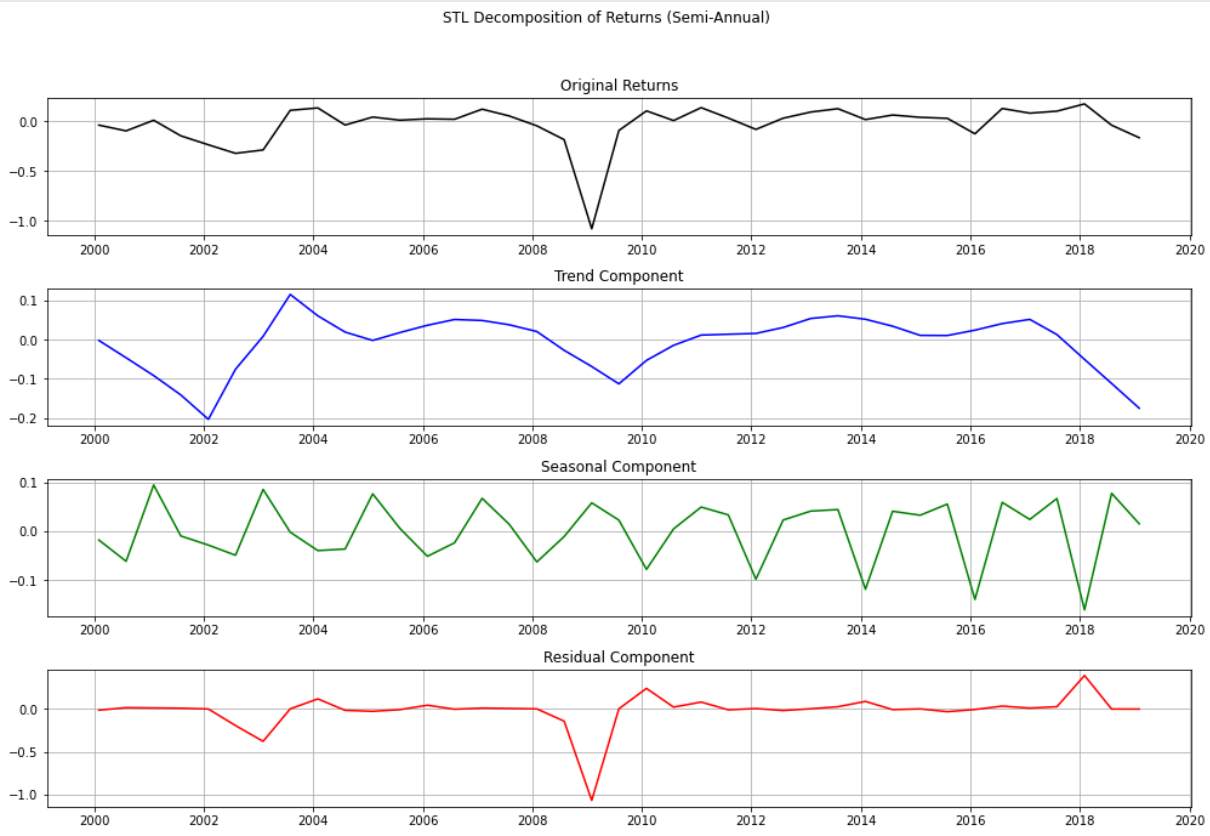
STL Decomposition of Returns





- The presence of clear trends and seasonality implies that structured, time-aware strategies are beneficial.
- The volatility in residuals suggests the importance of adaptive models that can handle sudden market changes.

STL Decomposition of Returns



Conclusion:

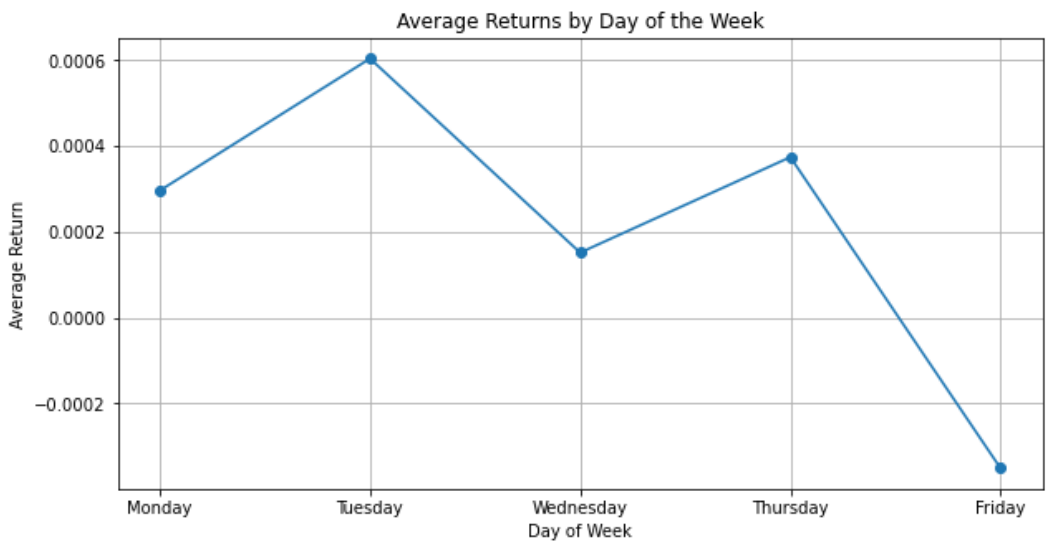
Based on the STL decomposition of returns periods:

- There are clear trends and consistent seasonality in timeframes.
- the residuals exhibit significant non-linear behavior, highlighting unpredictable market fluctuations.

This confirms that a hybrid SARIMA + SVM approach is highly effective.

SARIMA models the linear trends and seasonality captured across these timeframes, while SVM handles the complex, non-linear residuals, resulting in a more comprehensive and accurate forecasting model.

Average Returns by Day of the Week:

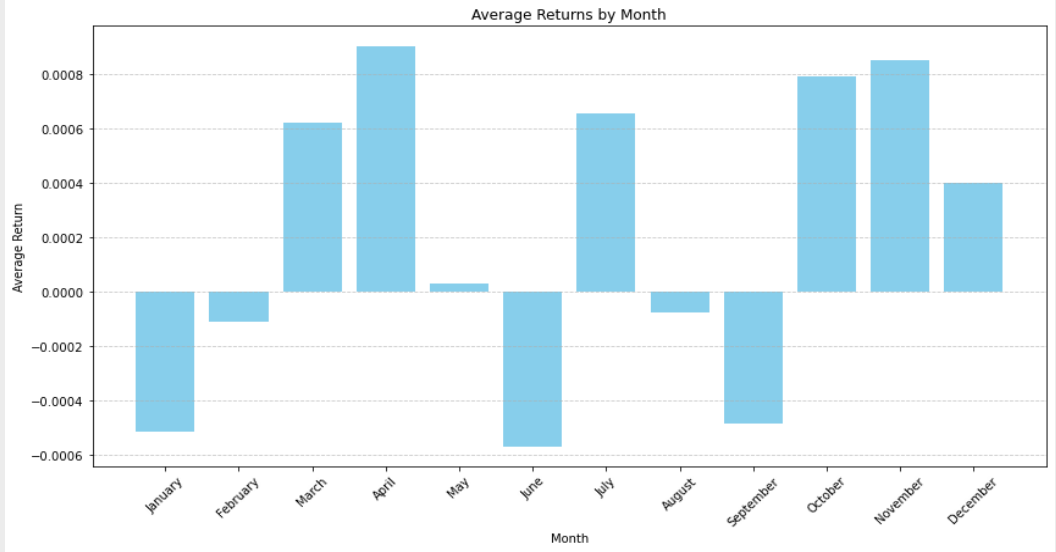


Observations:

- Returns are highest on Tuesdays and gradually decrease through the week.
- Friday shows a negative average return, indicating a potential end-of-week sell-off or market caution.

➢ This pattern suggests that there may be a mid-week peak in market optimism or trading activity, followed by caution as the week closes.

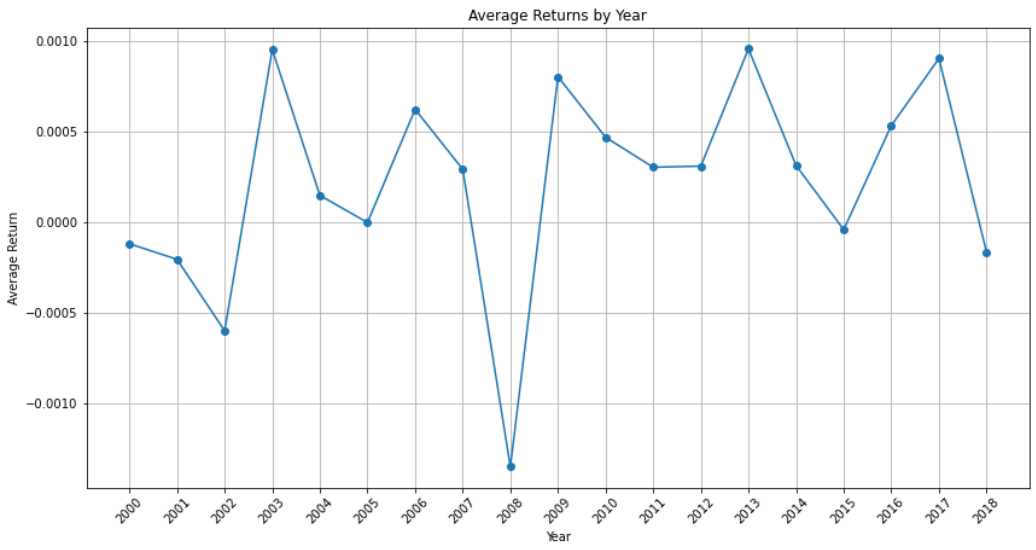
Average Returns by Month:



Observations:

- High Return Months: April, October and November. show these months are typically more favorable for the market.
- Low Return Months: June, August, and December have lower or negative returns, pointing to potential seasonal dips or volatility.

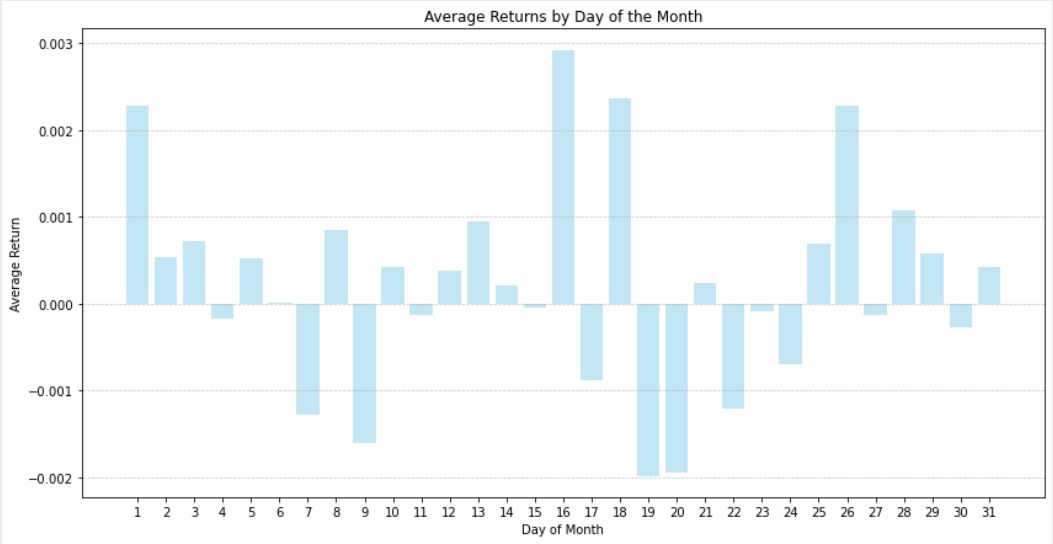
Average Returns by Year:



Observations:

- Yearly Volatility: Significant year-to-year volatility is observed, with sharp declines in years like 2008 and peaks in years like 2003 and 2013, reflecting economic cycles and major events.
- ➢ The cyclical pattern highlights the importance of a long-term perspective to navigate market fluctuations and manage the impact of economic shifts.

Average Returns by Day of the Month:



Observations:

- Higher returns on the 1st, 15th, and 26th; dips on the 10th, 20th, and 31st.
- ➢ Recognizing these patterns can aid in tactical decisions, such as timing entries and exits in the market around these recurring dates to optimize performance.

Introduction to Stationarity：

- Stationarity in time series :

statistical properties such as mean, variance, and autocorrelation structure do not change over time.

- A non-stationary time series:

trends, cycles, random walks, or other structures that change over time.

The Dickey-Fuller Test:

- The Dickey-Fuller test is used to determine whether a time series is stationary or not.
- This test is capable of revealing the presence of unit roots in the series, indicating non-stationarity.

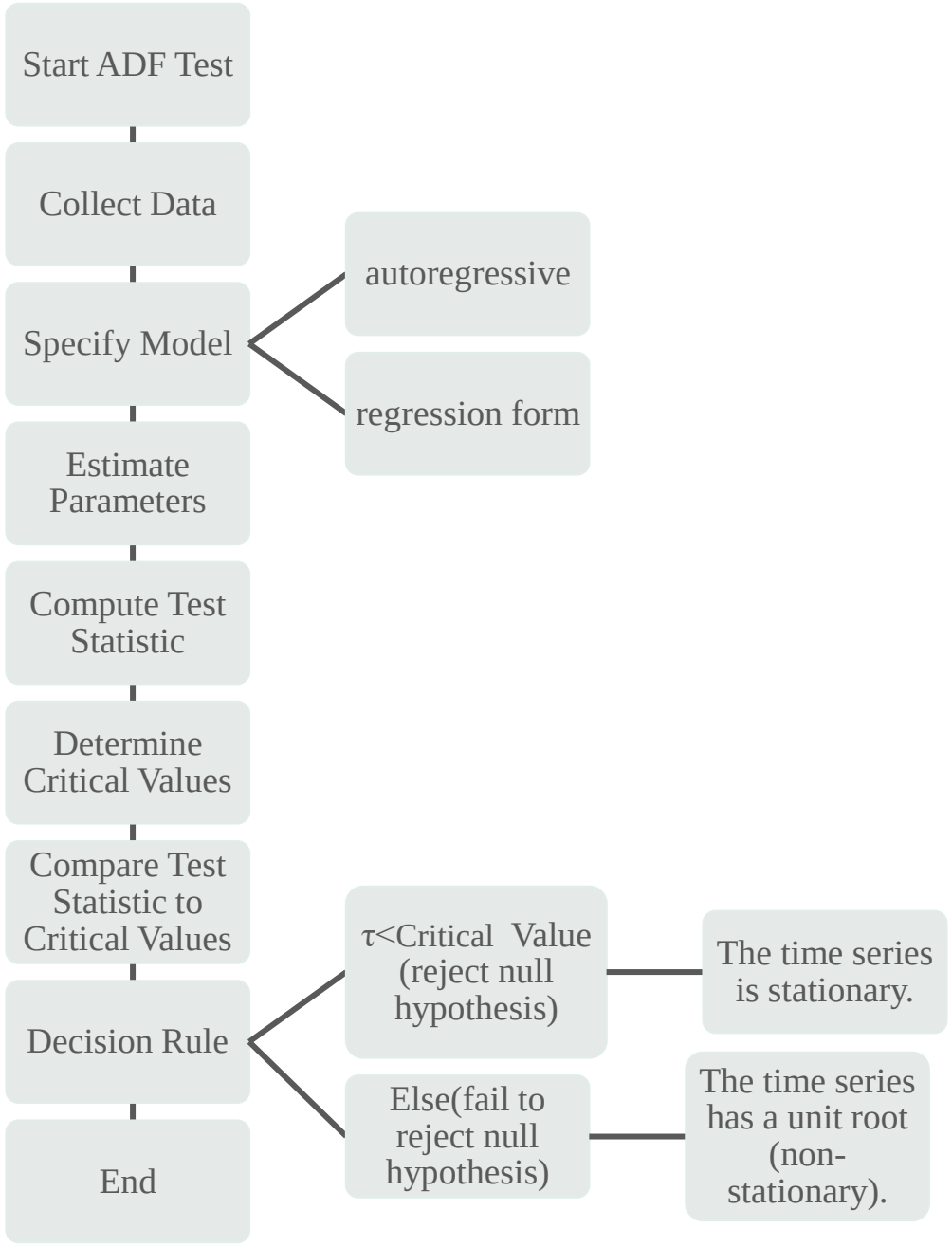
The Null and Alternative Hypotheses:：

- Null Hypothesis (H0): The time series has a unit root (i.e., it is non-stationary).
- Alternative Hypothesis (HA): The time series does not have a unit root (i.e., it is stationary).

The Dickey-Fuller Test Result:

- Test Statistic: -16.42285366209938
- P-value: 2.52e-29
- Critical Values: {'1%': -3.43, '5%': -2.86, '10%': -2.56}

➢ **The series is stationary.**



Kolmogorov-Smirnov Test:

Definition:

- The KS test is a nonparametric test used to compare a sample distribution with a reference probability distribution, or to compare two sample distributions.

Purpose:

- To determine if a sample comes from a population with a normal distribution.

Null Hypothesis (H_0):

- There is no difference between the observed distribution and the expected distribution.

Test Statistic:

- The maximum distance (D) between the empirical cumulative distribution function (ECDF) of the sample and the cumulative distribution function (CDF) of the reference distribution.

Interpretation:

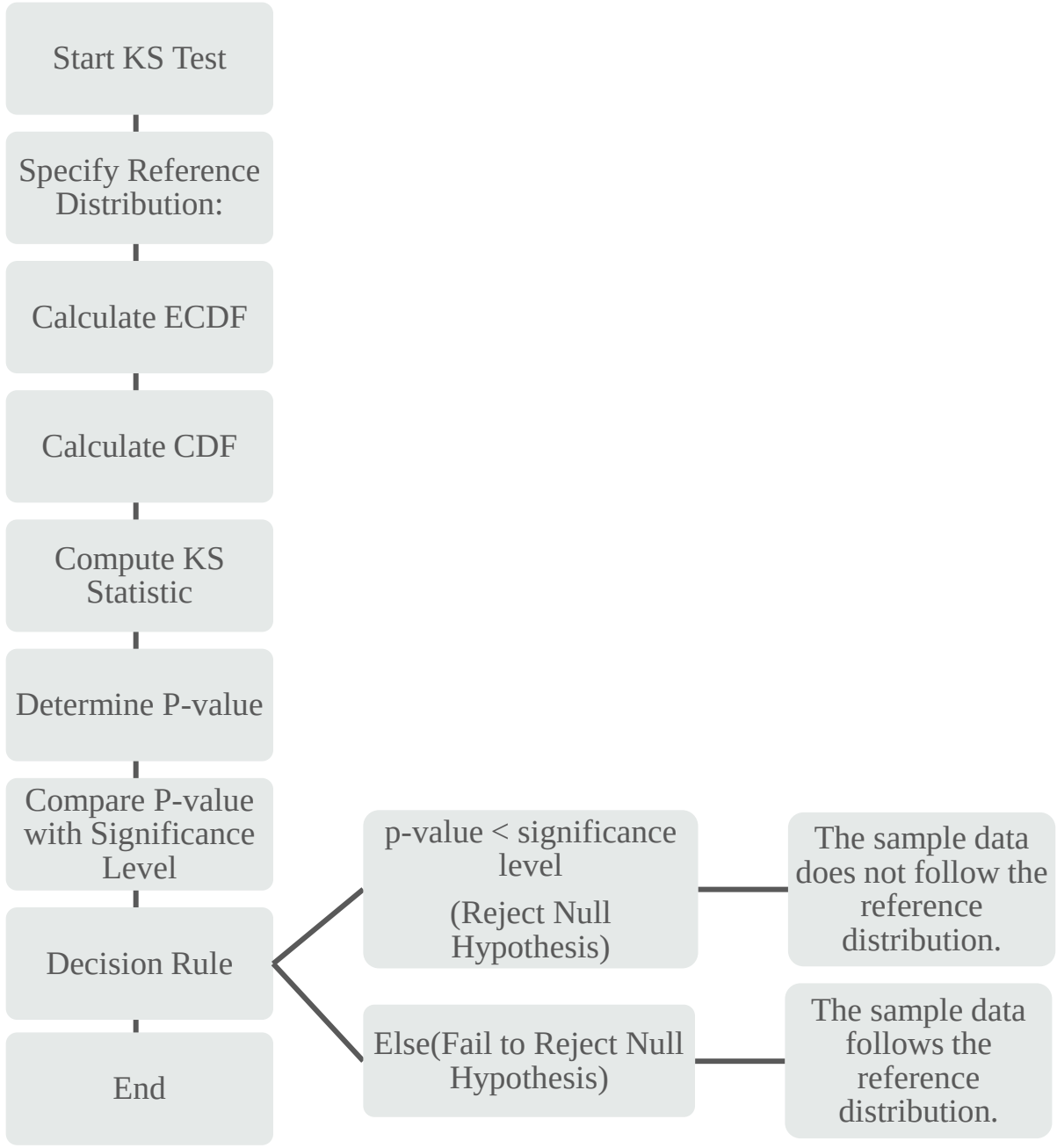
$$P - value(\text{e.g., } 0.05) > \alpha \rightarrow can\ not\ reject\ the\ null\ hypothesis$$

$$\mathcal{D} = \mathcal{S}\mathcal{u}\mathcal{p}_x |\, F_n\,(x) - F(x)|$$

- $\mathcal{D}$ is the Kolmogorov-Smirnov statistic.
- $\mathcal{S}\mathcal{u}\mathcal{p}_x$ denotes the supremum (the maximum value).
- $F_n\,(x)$ is the empirical cumulative distribution function (ECDF) of the sample.
- $F(x)$ is the cumulative distribution function (CDF) of the reference distribution.

The Kolmogorov-Smirnov Test Result:

- KS Statistic: 0.086
- p-value: 3.02e-31
- **The residuals are not normally distributed.**



Analyzing ADF and KS Tests for Residuals:

The Dickey-Fuller Test Result:

- ADF Statistic: -10.298
- p-value: 3.41e-18
- Critical Values:

1%: -3.43 -5%: -2.86 -10%: -2.567

The residuals are stationary (reject the null hypothesis).

The Kolmogorov-Smirnov Test Result:

- KS Statistic: 0.2128
- p-value: 1.4473e-117

➢ **The residuals do not follow a normal distribution** (reject the null hypothesis).

Residuals Analysis:
Stationary residuals confirm that the model captures all trends and seasonality, leaving only random noise, which is essential for reliable forecasting.
residuals not following a normal distribution indicate that traditional linear models like ARIMA or SARIMA may not fully address the data's complexities.
Solution:
A hybrid model combining SARIMA (for linear trends and seasonality) and SVM (for non-linear deviations) is used to capture both linear and non-linear patterns, enhancing forecasting accuracy and robustness.

Introducing Models:

Autocorrelation Function (ACF):

Definition:

- ACF is a statistical tool used to identify repeating patterns or cycles in time series data.
- measured by the correlation between observations separated by time units.
- Use Case: Identifying the order of moving average (MA) models in time series analysis.

Formula:

$$\rho_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X}_t)(X_{t+k} - \bar{X}_t)}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

- ρ_k: Autocorrelation at lag k
- X_t : Value of the time series at time t
- $\bar{X}$: Mean of the time series
- n: Number of observations

The partial autocorrelation function (PACF):

Definition:

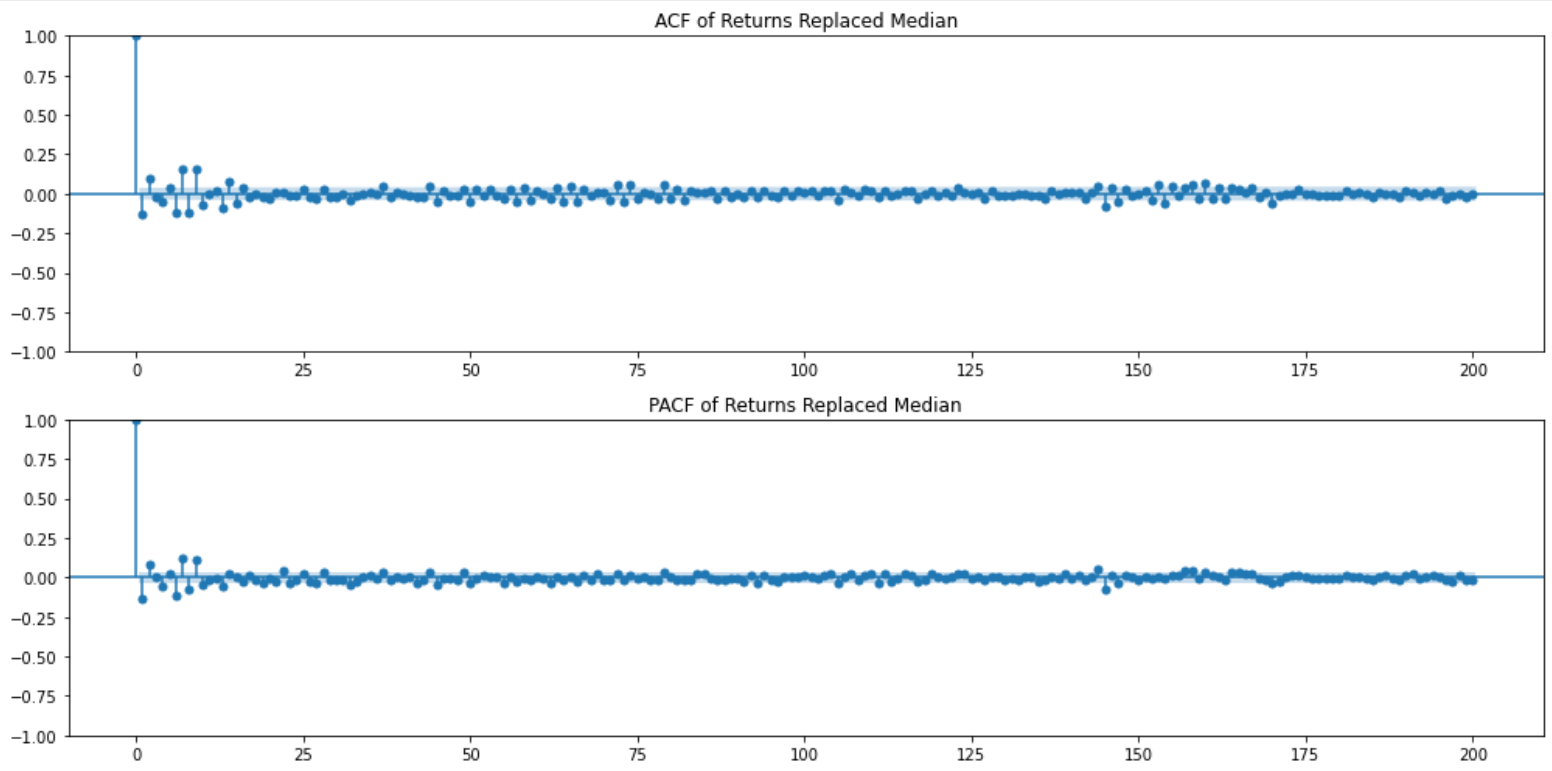
- PACF measures the correlation between observations of a time series separated by K time units, after removing the effects of correlations at shorter lags.
- It helps to identify the direct relationship between observations at lag k.
- Use Case: Identifying the order of autoregressive (AR) models in time series analysis.

Formula:

$$\phi_{kk} = \frac{\rho_k - \sum_{i=1}^{k-1}\phi_{ki}\rho_{k-i}}{1 - \sum_{i=1}^{k-1}\phi_{ki}\rho_i}$$

- ϕ_{kk}: Partial autocorrelation at lag k.
- ρ_k: Autocorrelation at lag k
- ϕ_{ki}: Partial autocorrelation at lag k given i

ACF and PACF:



Result:

The ACF and PACF plots show immediate autocorrelation at lag 1, suggesting AR(1) could capture linear patterns.
However, to address non-linear relationships, a hybrid model like SARIMA + SVM is more suitable.

Introducing Models:

SARIMA (Seasonal ARIMA) Model

Definition:

- SARIMA stands for Seasonal AutoRegressive Integrated Moving Average.
- Extends ARIMA by incorporating seasonal components, capturing trends, seasonality, and noise.

Formula:

$$\mathbf{SARIMA(p, d, q) \times (P, D, Q, s)}$$

- p, d, q: Non-seasonal autoregressive, differencing, and moving average orders.
- P, D, Q, s: Seasonal autoregressive, differencing, moving average orders, and season length.

Components:

- **AR** (AutoRegressive): Relates current values to past values.
- **MA** (Moving Average): Models the error as a combination of past errors.
- **Seasonal Components**: Address repeating patterns at regular intervals.

Support Vector Machine (SVM) for Regression

Definition:

- SVM is a machine learning method that captures complex, non-linear relationships in data.
- Effective for regression tasks where traditional models may fail to capture non-linear dynamics.

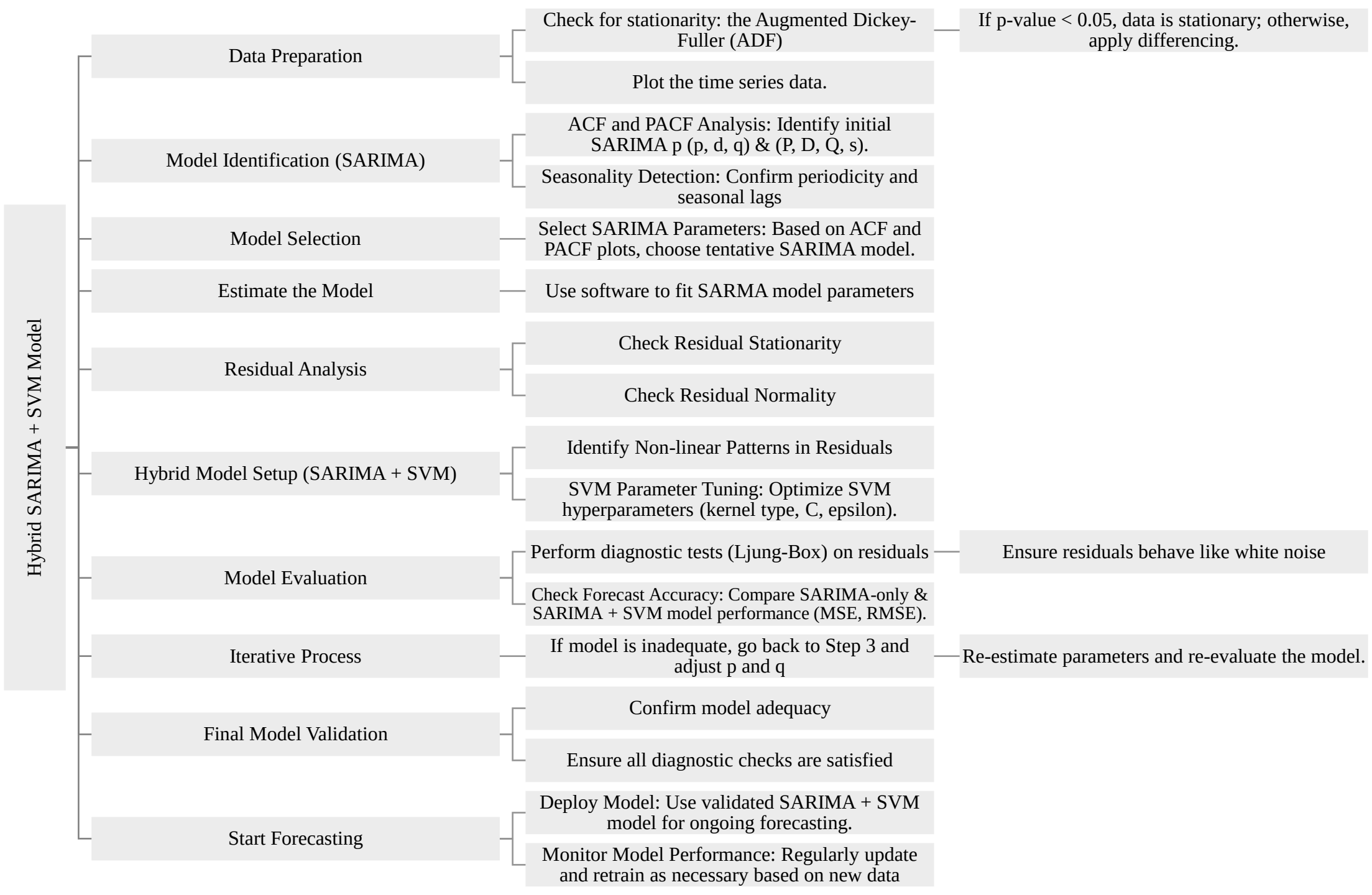
Key Features:

- Kernel Trick: Transforms data into higher dimensions, making it easier to find patterns.
- Support Vectors: Data points that define the decision boundary.
- Hyperplane: A line or plane that best fits the data in high-dimensional space.

Advantages:

- Flexibility: Handles non-linear relationships well.
- Robustness: Less sensitive to outliers in the data set.
- Versatility: Applicable to both classification and regression tasks.

Hybrid SARIMA + SVM Model Flowchart:



Hybrid SARIMA + SVR Model Forecast:

To forecast stock returns by combining: SARIMA and SVR modeling techniques:

- SARIMA: Captures trends and seasonality.
- SVR: Addresses residual non-linear patterns.

Insights:

- Linear Trends: Managed by SARIMA.
- Non-Linear Residuals: Corrected by SVR.
- Improvement: Enhanced accuracy by addressing both patterns.

Implications:

- Suitable for complex data with both predictable and irregular behaviors.
- Balances volatility smoothing and detailed prediction.

Model Parameters:

➢ SARIMA Model:

Order: (3, 0, 5)
Seasonal Order: (1, 1, 0, 10)
Optimization Method: L-BFGS-B
Maximum Iterations: 300

➢ SVR Model:

Optimized Parameters:

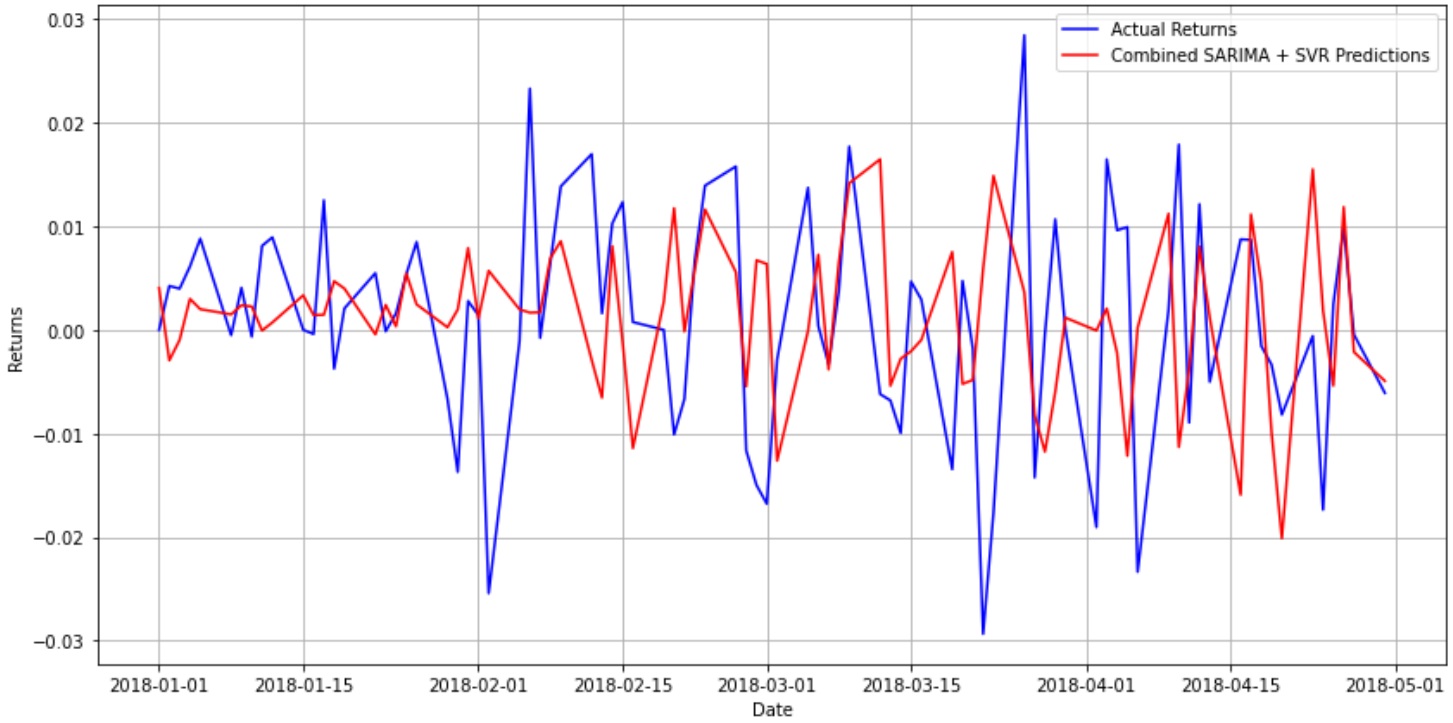
✓ C(Regularization Parameter): (0.1, 10)
✓ Epsilon(Insensitivity Zone): (0.01, 0.5)
✓ Kernel: RBF (Radial Basis Function)
✓ Model Selection Process: Used with 200 iterations 5-fold cross-validation for SVR hyperparameter optimization.

Hybrid SARIMA + SVR Model Forecast:

Forecast Performance (Model Evaluation):

- Metrics: MSE:0.00016 MAE: 0.0095 RMSE:0.01278

- The SARIMA + SVR model accurately predicts return trends, with predictions closely aligning with actual data.
- It demonstrates robust predictive accuracy, as indicated by the MSE, MAE, and RMSE metrics, particularly during stable market periods.
- While slight deviations are observed during volatile periods, suggesting room for refinement, the model remains a reliable baseline for forecasting market returns.
- Overall, it offers significant value in practical applications, effectively capturing market trends and providing actionable insights.



Model Validation:

Cross-Validation MSE Scores:

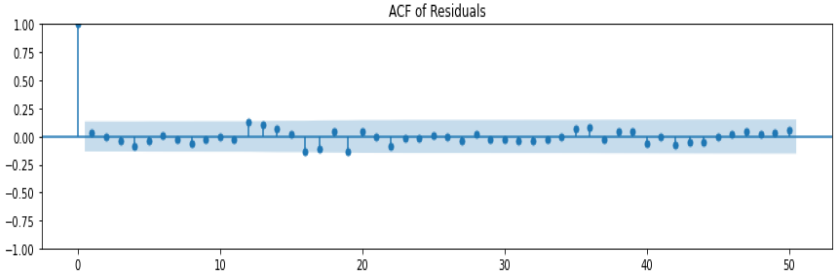
- The Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) indicate reasonable predictive accuracy, although some discrepancies exist, especially during volatile periods.

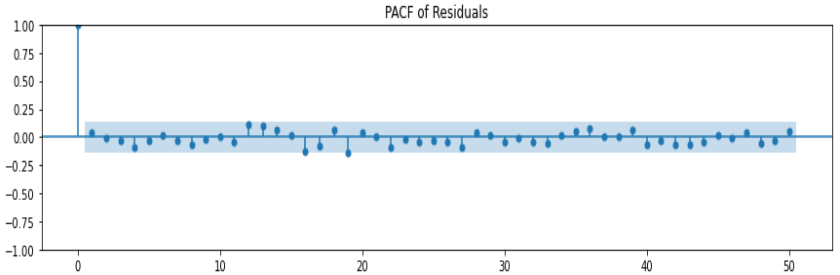
Ljung-Box Test:

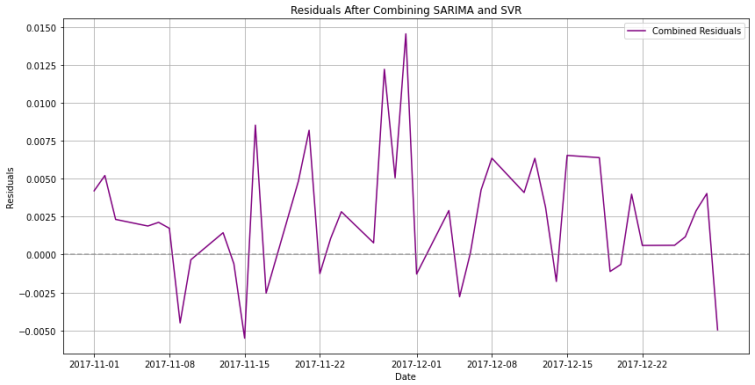
- The Ljung-Box test checks for the presence of autocorrelation in the residuals of the model.
- Purpose: To validate that residuals behave like white noise (i.e., no autocorrelation).
- Result: The test results show a p-value of 0.431666, suggesting that there is no significant autocorrelation left in the residuals, which confirms the model's adequacy.

Kolmogorov-Smirnov (KS) Test for Normality of Residuals::

- The KS Statistic is 0.0867 with a p-value of 0.8669, indicating that the residuals are normally distributed, supporting the assumptions required for effective time series modeling.







Residual Analysis:

ACF and PACF of Residuals:

- The ACF and PACF plots show no significant autocorrelation in the residuals after lag 1, suggesting that the residuals behave like white noise.
- This supports that the model is well-specified and effective in capturing the patterns within the data.
- **Residuals (eta) Plot:**
- Residuals mostly resemble white noise, fluctuating around zero without a clear pattern, indicating effective model performance. Some variance inconsistencies and spikes are observed, particularly during high volatility periods.
- Statistical tests, including the Ljung-Box and Kolmogorov-Smirnov tests, confirm no significant autocorrelations beyond lag 1, indicating that the model effectively captures the data's patterns, leaving only random noise in the residuals.
- Despite exhibiting some white noise properties, The model's residuals show variance inconsistencies, especially during high volatility periods, indicating that while it performs well overall, further refinements may be needed for improved stability.

Conclusion:

Model Effectiveness:

- The SARIMA + SVR hybrid model effectively captures key patterns in stock returns, as indicated by low error metrics (MSE, MAE, RMSE).
- Residual analysis shows that the model successfully removes autocorrelations, resembling white noise, although some variance inconsistencies are observed.

Identified Limitations:

- The model faces challenges in accurately capturing extreme market volatility, suggesting that the current approach may not fully address sudden shifts in market behavior.

Opportunities for Improvement:

- Advanced Volatility Modeling:
  Consider exploring models like GARCH to better handle periods of high volatility and refine the model's response to extreme market conditions.
- Integration of Non-Linear Methods:
  Further enhance the hybrid approach by incorporating more sophisticated non-linear techniques, potentially improving the model’s ability to capture complex market behaviors.

Future Recommendations:

- Iterative Model Refinement:
  Regularly revisit the model's performance with new data to continuously improve its accuracy and robustness.
- Enhanced Validation Practices:
  Implement more comprehensive validation techniques, including stress testing under various market scenarios, to ensure the model's adaptability and reliability.

References:

- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.
- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts. Retrieved from https://otexts.com/fpp3/
- Ljung, G. M., & Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models. Biometrika, 65(2), 297-303.
- Wei, W. W. S. (2006). Time Series Analysis: Univariate and Multivariate Methods. Pearson Education.
- Lv, P.; Wu, Q.; Xu, J.; Shu Y. Stock Index Prediction Based on Time Series Decomposition and Hybrid Model. Entropy 2022, 24, 146. https://doi.org/10.3390/e24020146
- Journal of Statistical and Econometric Methods, vol.4, no.4, 2015, 41-53 , ISSN: 1792-6602 (print), 1792-6939 (online) , Scienpress Ltd, 2015