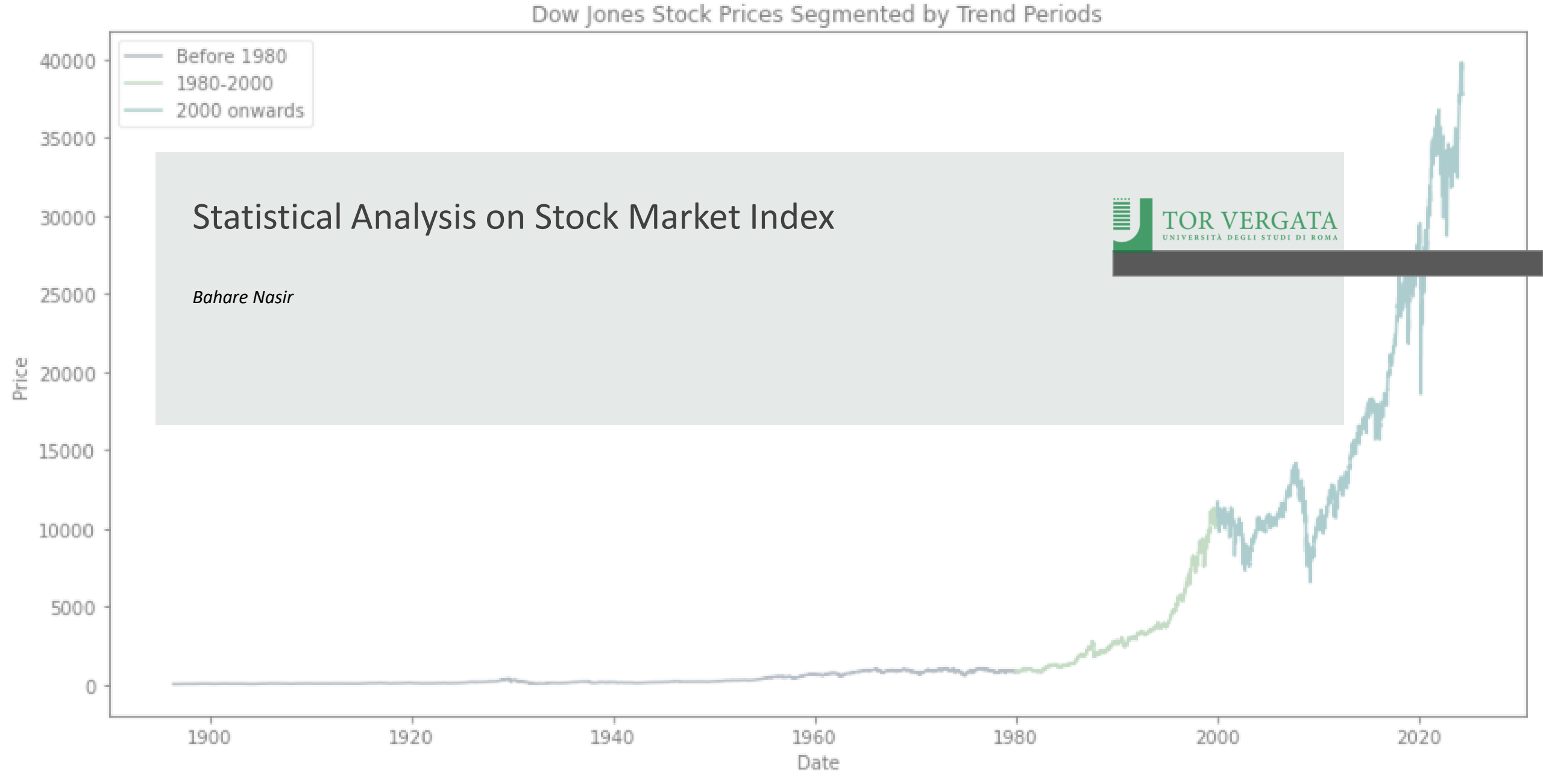


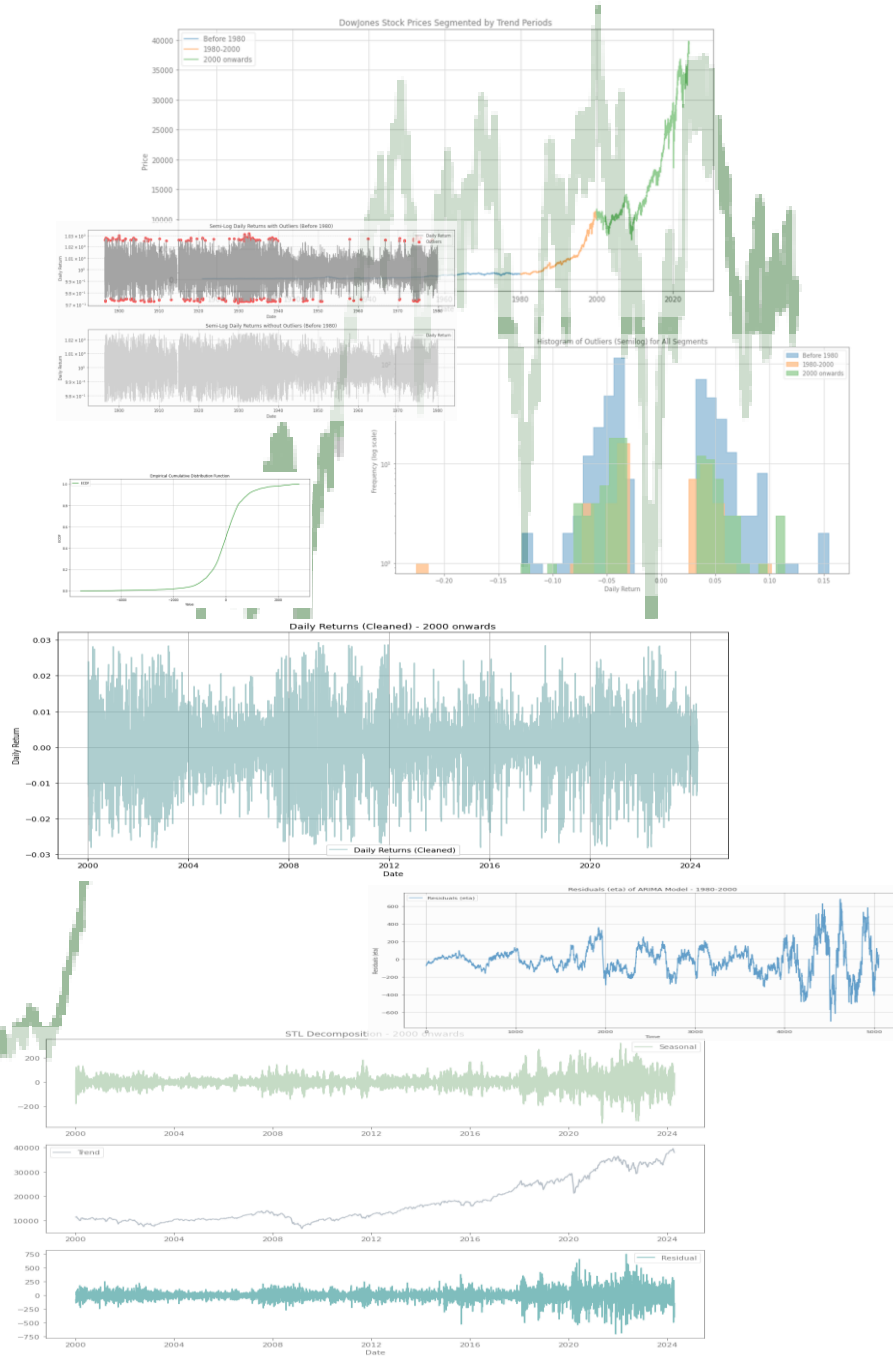
Statistical Analysis on Stock Market Index

Bahare Nasir

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Context:

- Introduction
- Suggested research methodology
- Data Preparation and Segmentation
- Outlier Detection and Analysis
- Temporal Analysis of Outliers
- STL Decomposition
- The Dickey-Fuller Test
- Kolmogorov-Smirnov Test
- Quantile Normalization of Residuals
- Introducing Models
- Conclusion
- Future Work Suggestions – Forecasting
- 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

Indices:

- Dow Jones Industrial Average measures overall market performance

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 examines sequential data points recorded over time to extract insights, identify patterns, and make predictions.

Key Techniques:

- ARIMA:
  - Models trends and seasonal variations
  - Useful for forecasting
- GARCH:
  - Models cyclical patterns and volatility
  - Helps in risk assessment

Linking Time Series Analysis and Stock Market Dynamics:

Time series analysis is essential in stock market research for uncovering patterns and forecasting future movements.

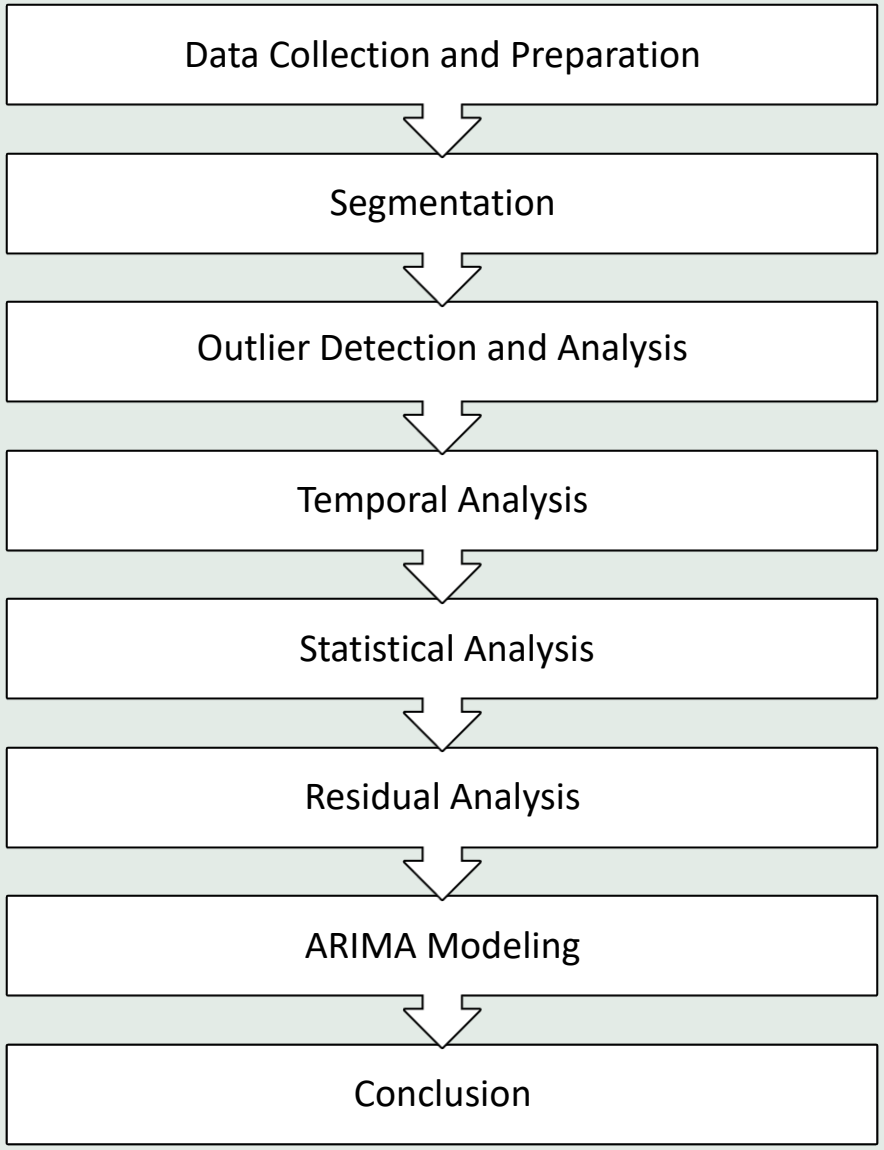
By applying techniques such as ARIMA to stock market indices like the Dow Jones, analysts can:

- Models trends and seasonal effects
- Captures market volatility

Benefits:

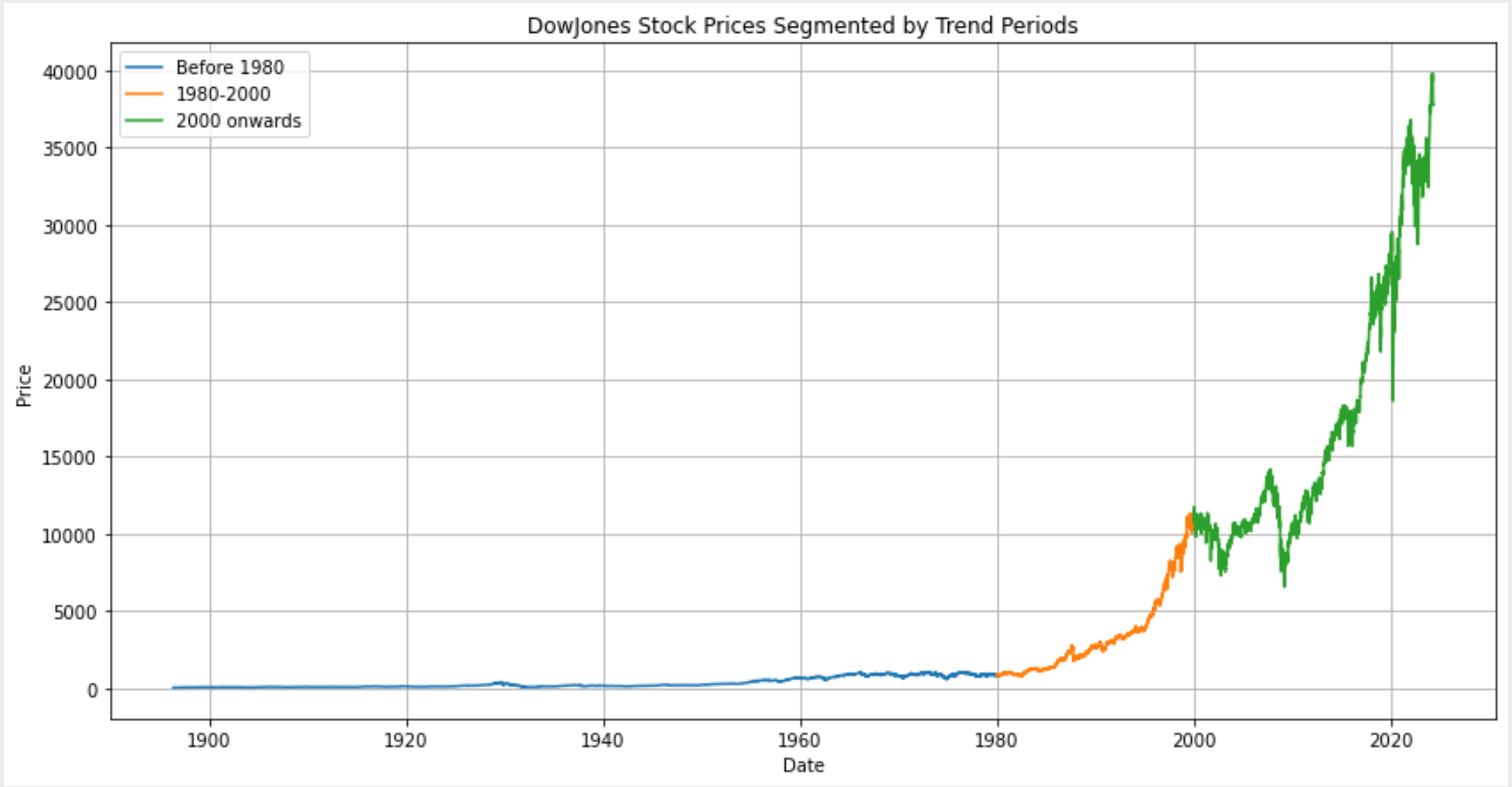
- Informed investment decisions
- Enhanced understanding of market dynamics

Suggested research methodology:



Data Preparation and Segmentation:

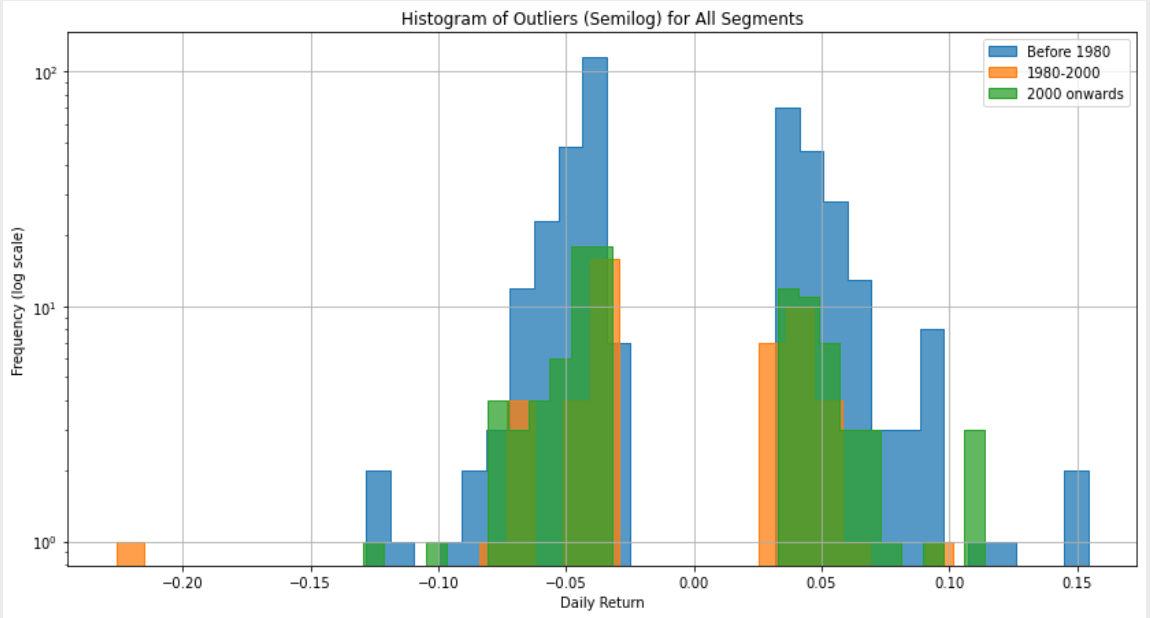
Analyze different market behaviors and trends over time:

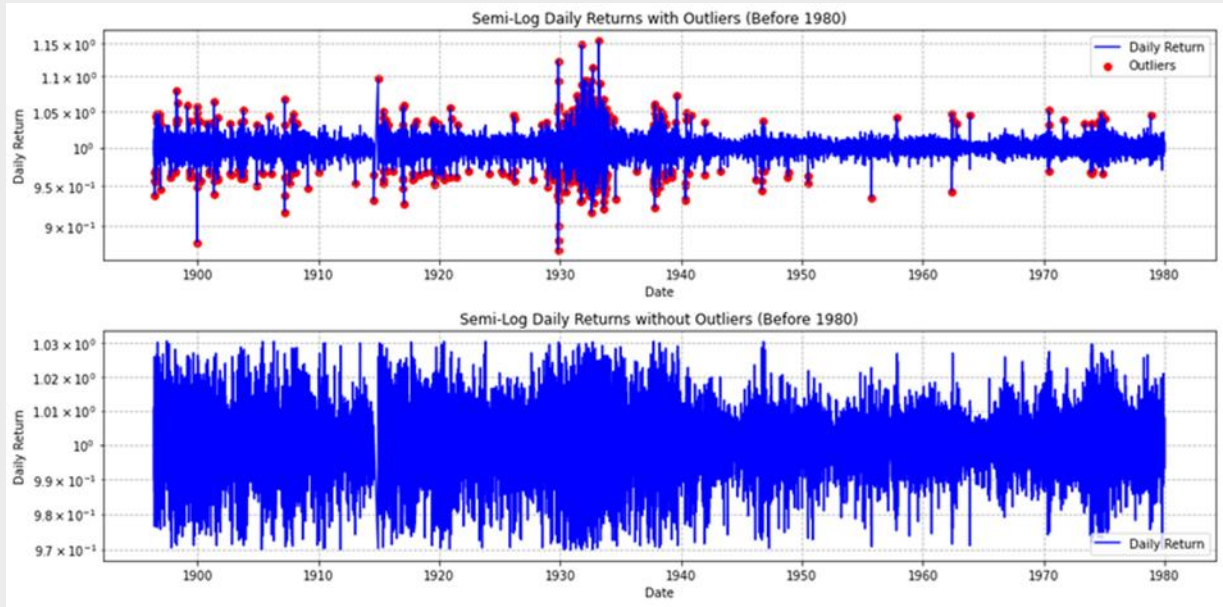


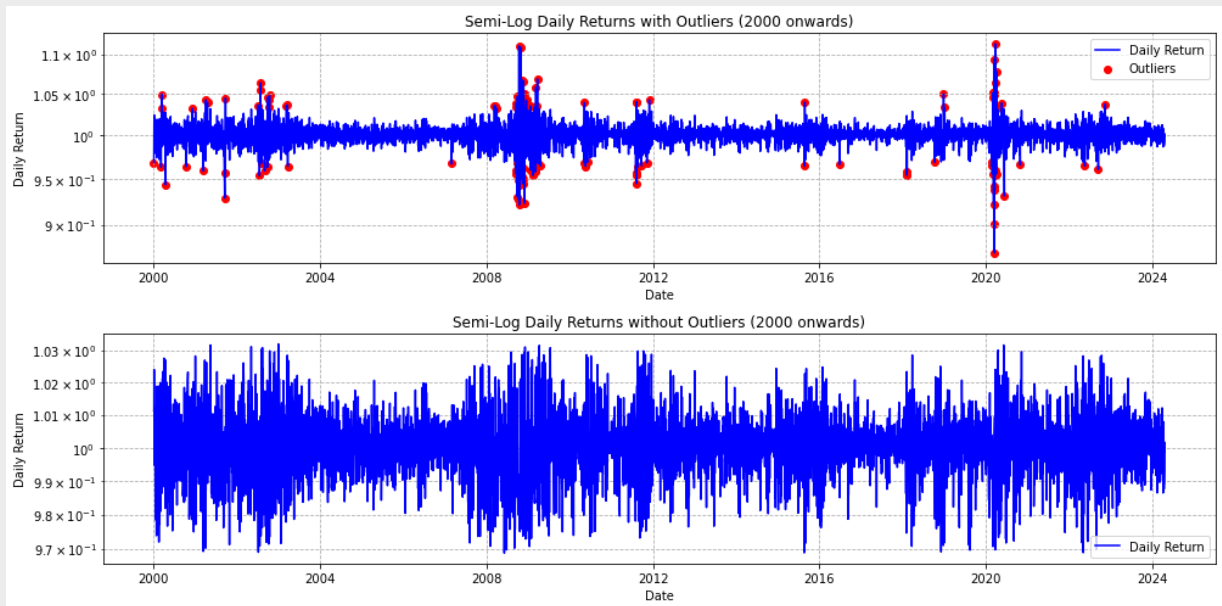
Outlier Detection and Analysis:

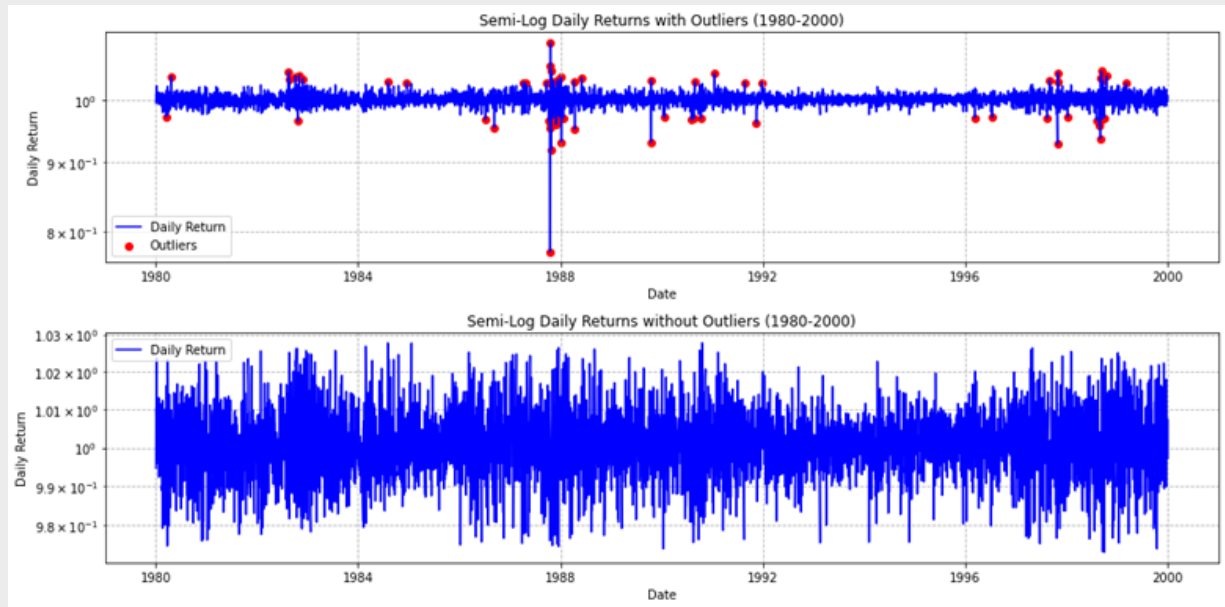
Outlier Detection using Z-Score:

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Number of outliers: 390 • Total days: 21949 • Percentage of outliers: 1.78%	• Number of outliers: 49 • Total days: 5055 • Percentage of outliers: 0.97%	• Number of outliers: 96 • Total days: 6111 • Percentage of outliers: 1.57%




Semi-Log Daily Returns with Outliers (Before 1980)
Daily Return
Outliers
Daily Return
Date
Semi-Log Daily Returns without Outliers (Before 1980)
Daily Return
Date

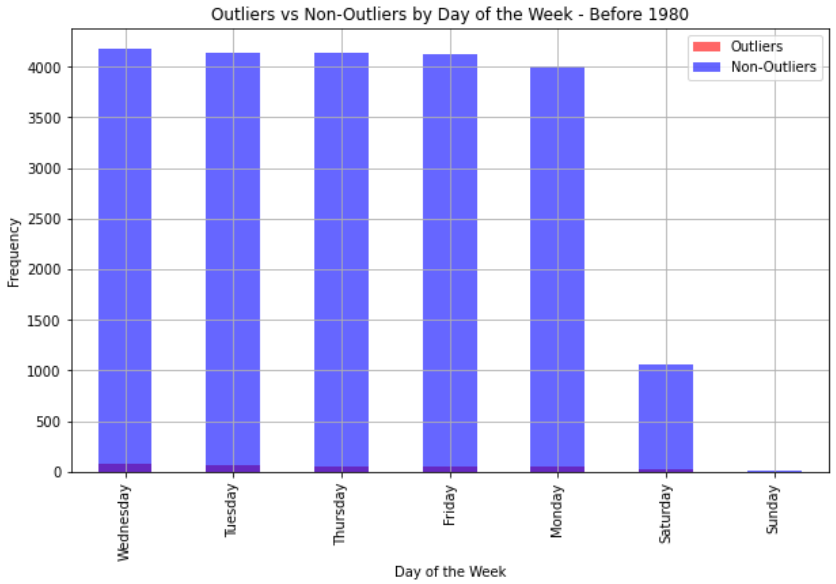

Semi-Log Daily Returns with Outliers (2000 onwards)
Daily Return
Outliers
Daily Return
Date
Semi-Log Daily Returns without Outliers (2000 onwards)
Daily Return
Date

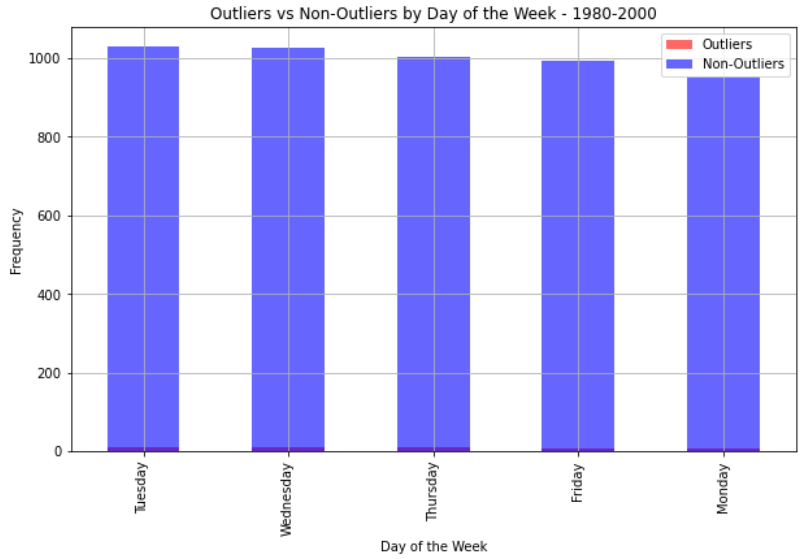

Semi-Log Daily Returns with Outliers (1980-2000)
Daily Return
Daily Return
Outliers
Date
Semi-Log Daily Returns without Outliers (1980-2000)
Daily Return
Daily Return
Date

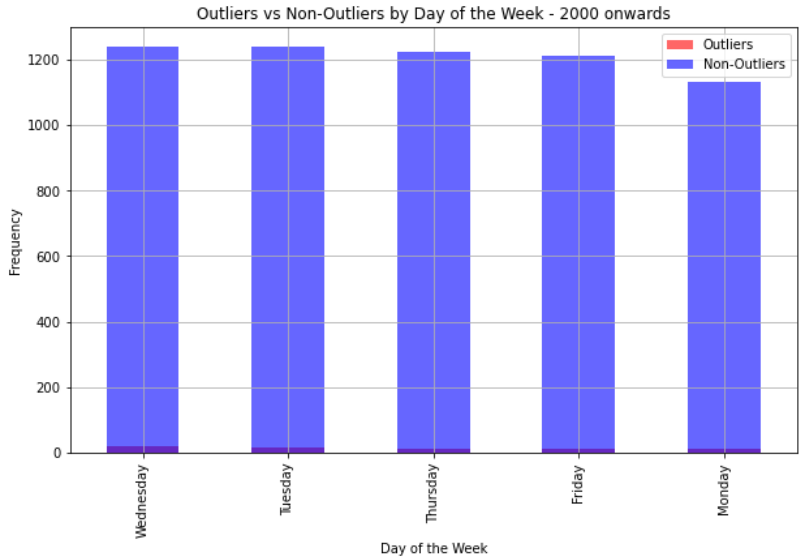
Temporal Analysis of Outliers:

Examining the Outliers by Day:

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Outliers and non-outliers show a similar distribution across weekdays. • Slightly fewer data points on Saturdays.	• Consistent distribution across weekdays. • Minimal outlier presence on weekends.	• Similar weekday distribution. • Non-outliers dominate Saturdays.







Statistical Summary:

With Outliers:

- Mean: 0.0002006120043892403
- Median: 0.00045372050816694376
- Std Dev: 0.009786779944221355
- Skewness: -0.08680493685088284
- Kurtosis: 2.3858778313292817

Without Outliers:

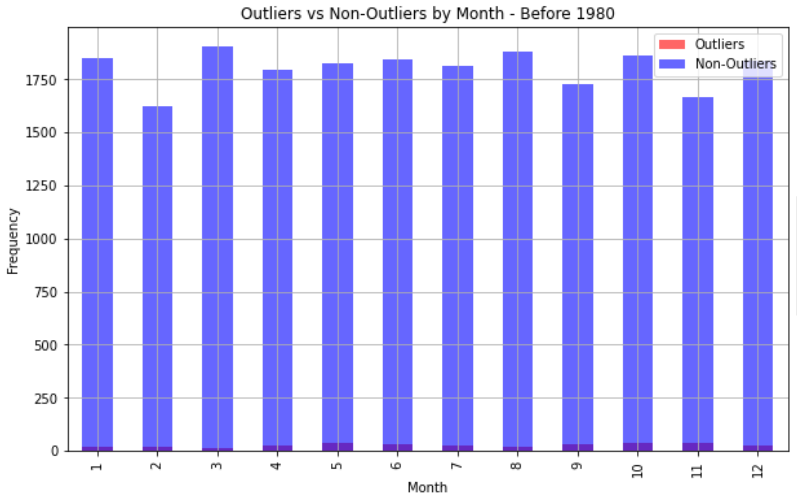
- Mean: 0.00024207818346210657
- Median: 0.00046965780105079347
- Std Dev: 0.008929338671849692
- Skewness: -0.13288713003128094
- Kurtosis: 0.8885712442784954

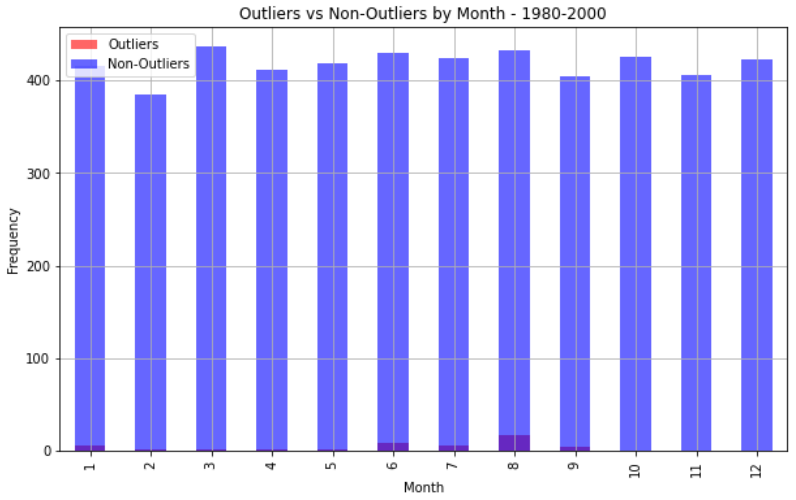
Conclusion:
The removal of outliers slightly increases the mean and median while reducing the standard deviation, skewness, and kurtosis, indicating a more stable and less extreme distribution of returns.

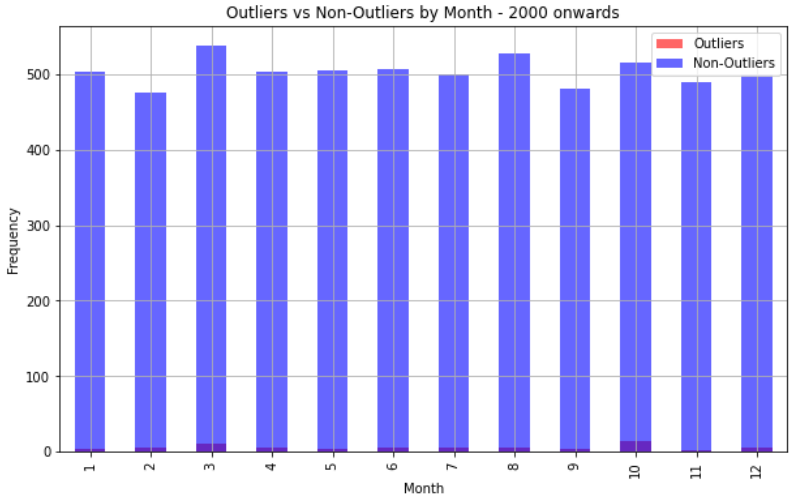
Temporal Analysis of Outliers:

Examining the Outliers by Month:

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Even distribution of non-outliers through out the year. • Outliers are sparse but appear consistently across all months.	• Similar to the previous period, with an even distribution. • Outliers are rare, with slight peaks in some months.	• Consistent monthly distribution. • Outliers occur sporadically throughout the year.







Statistical Summary:

With Outliers:

Mean: 0.000561613480491588

Median: 0.0005327934359848907

Std Dev: 0.008970794891200454

Skewness: -0.001583472030223251

Kurtosis: 1.0320307054299596

Without Outliers:

Mean: 0.0005737643109942507

Median: 0.0005402339212878804

Std Dev: 0.008539430530476962

Skewness: 0.008509281734478317

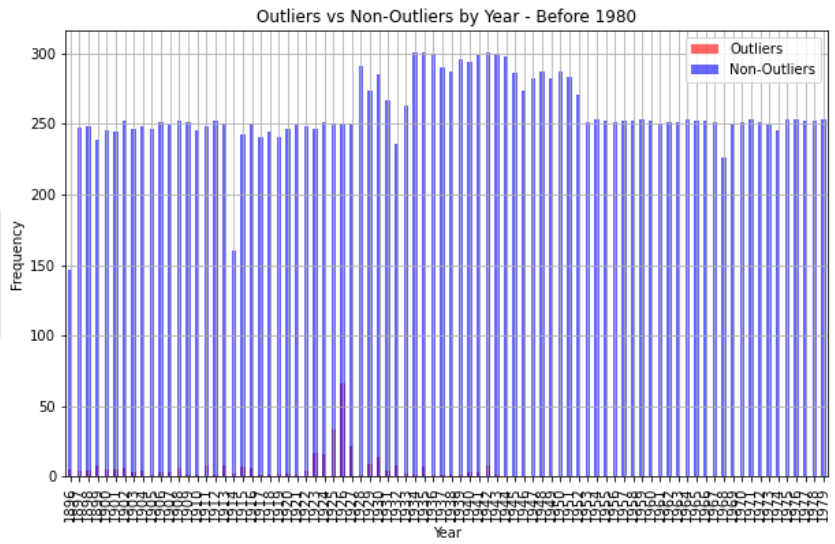
Kurtosis: 0.5013315341084388

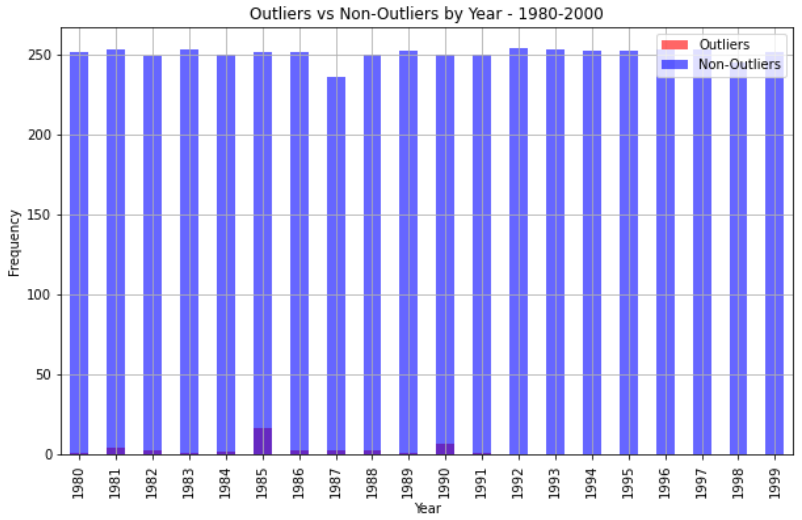
Conclusion:
Removing outliers results in higher mean and median, and reduces the standard deviation, skewness, and kurtosis, suggesting a more normal distribution of returns.

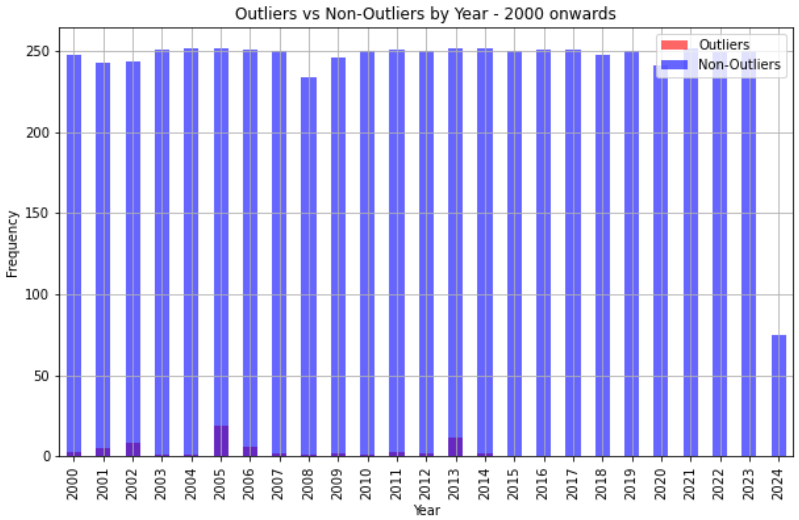
Temporal Analysis of Outliers:

Examining the Outliers by Year:

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Steady increase in non-outlier frequency over the years. • Significant outlier spikes during major market events (e.g., Great Depression).	• Uniform non-outlier distribution with minor dips. • Noticeable outlier peaks during market crashes (e.g., Black Monday 1987).	• Continuous increase in non-outliers. • Outliers more frequent during financial crises (e.g., 2008 crisis, COVID-19 pandemic).







Statistical Summary:

With Outliers:

- Mean: 0.000247895081587936
- Median: 0.00046857553038770483
- Std Dev: 0.010119739204024062
- Skewness: -0.03882570290945209
- Kurtosis: 2.1476866926536102
- **Without Outliers:**
- Mean: 0.00027789536495140623
- Median: 0.00048354028857677 81
- Std Dev: 0.0094280732318196
- Skewness: -0.13231775146276406
- Kurtosis: 0.8300442152569953

Conclusion:
Similar to previous segments, removing outliers increases the mean and median while reducing the standard deviation, skewness, and kurtosis. This pattern highlights that outliers significantly influence the statistical properties of the data.

Statistical Analysis of Cleaned Data:

Before 1980 Segment:

- Mean: 306.8132
- Median: 152.3
- Std Dev: 308.1563
- Skewness: 1.0502
- Kurtosis: -0.4987

Conclusion:

- Significant increase in mean and median close prices after removing outliers.
- Positive skewness and slightly negative kurtosis indicate more frequent moderate positive returns.

1980-2000 Segment:

- Mean: 3401.8538
- Median: 2626.55
- Std Dev: 2664.1004
- Skewness: 1.3394
- Kurtosis: 0.9075

Conclusion:

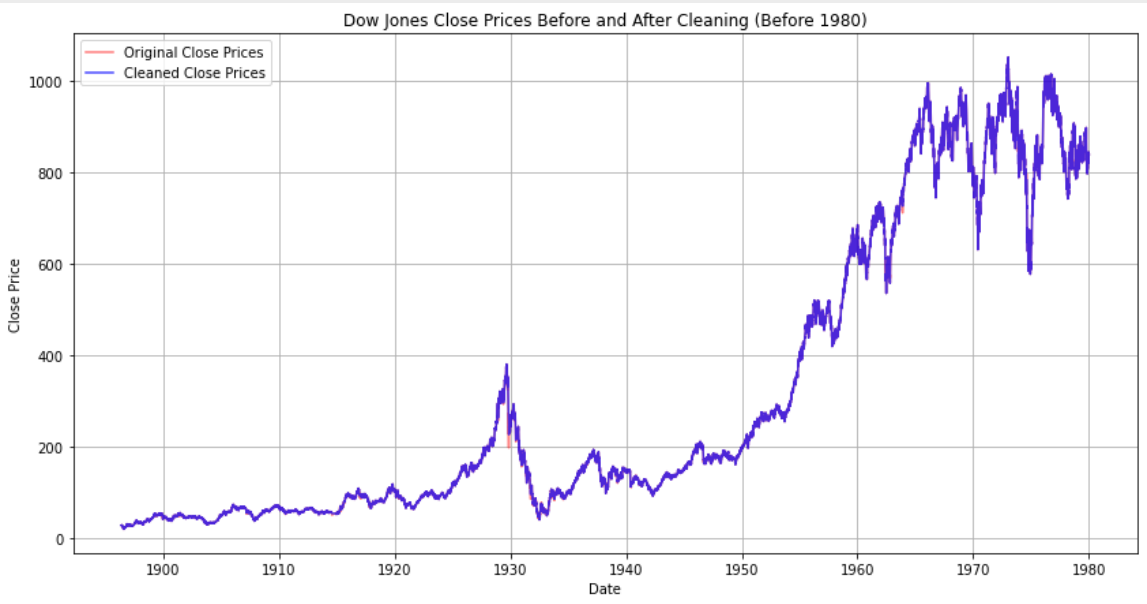
- mean and median close prices with increased volatility.
- High positive skewness indicates a long right tail.
- Positive kurtosis reflects frequent extreme returns.

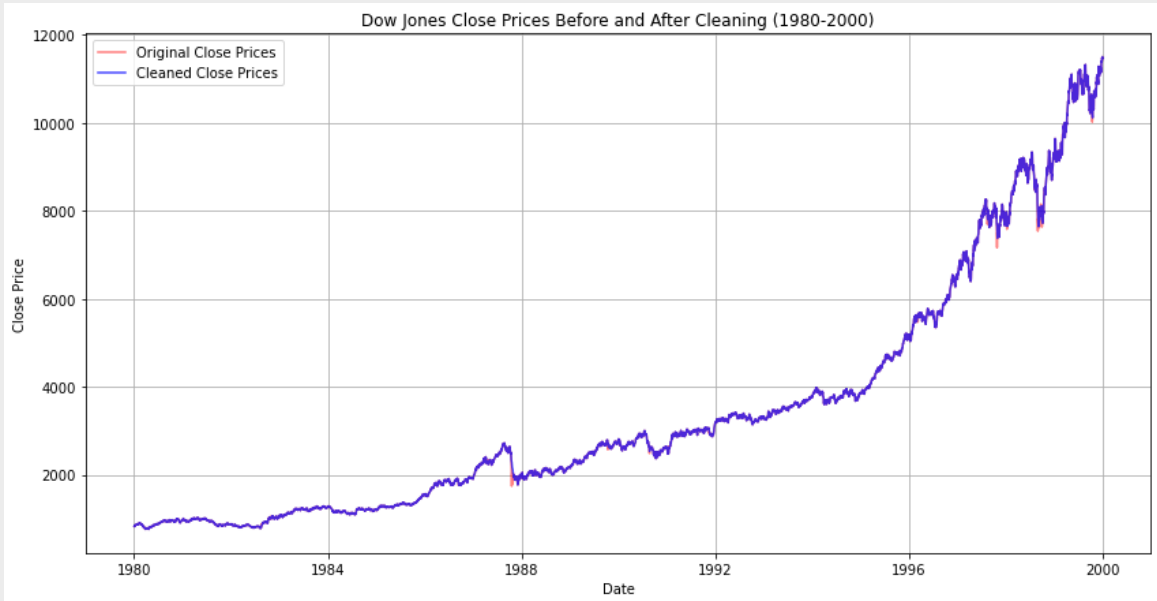
2000 onwards Segment:

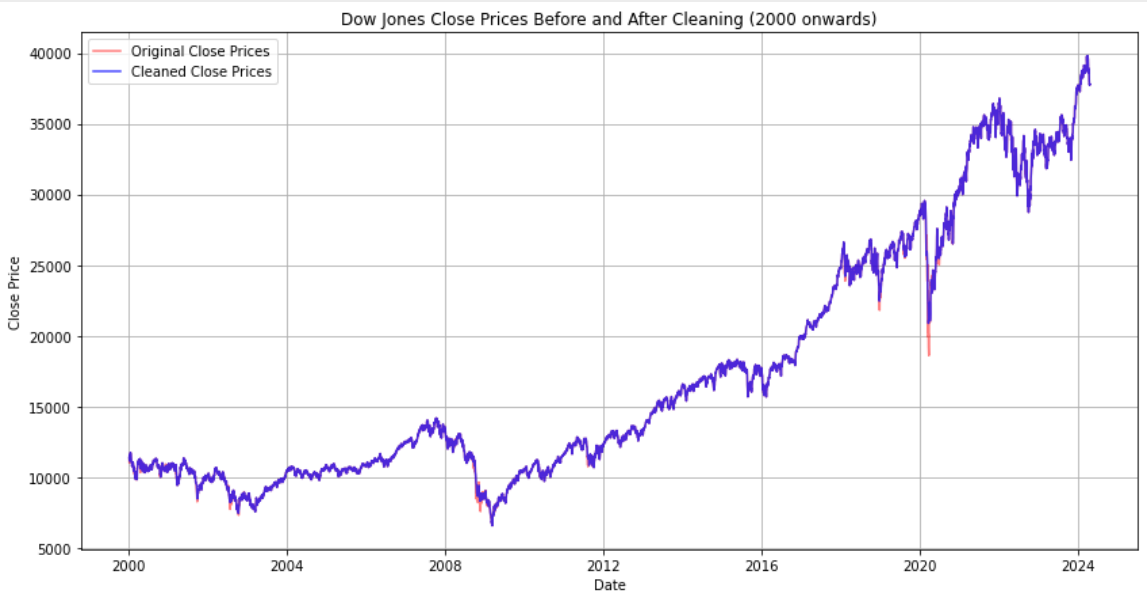
- Mean: 17298.6278
- Median: 13176.22
- Std Dev: 8602.1312
- Skewness: 0.9520
- Kurtosis: -0.4175

Conclusion:

- Significantly higher mean and median close prices.
- High volatility.
- Positive skewness and slightly negative kurtosis suggest frequent moderate positive returns.


Dow Jones Close Prices Before and After Cleaning (Before 1980)
Original Close Prices
Cleaned Close Prices
Close Price
Date
0
200
400
600
800
1000
1900
1910
1920
1930
1940
1950
1960
1970
1980

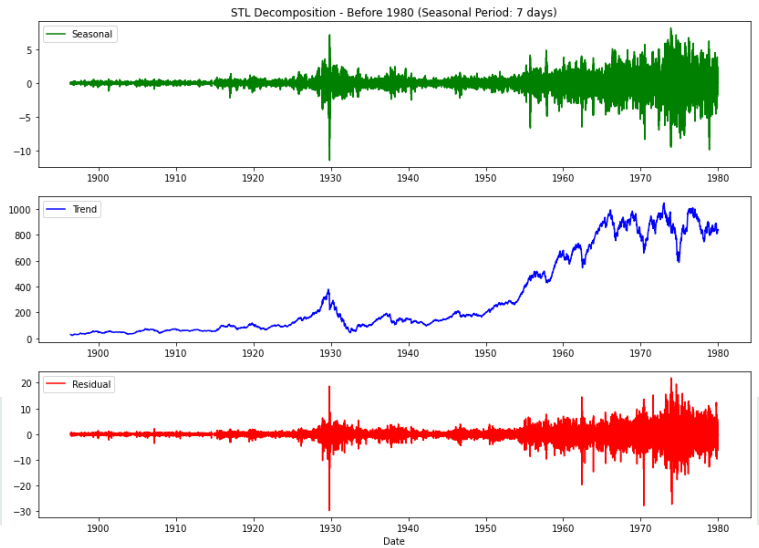

Dow Jones Close Prices Before and After Cleaning (1980-2000)
Original Close Prices
Cleaned Close Prices
Close Price
Date
2000
4000
6000
8000
10000
12000
1980
1984
1988
1992
1996
2000

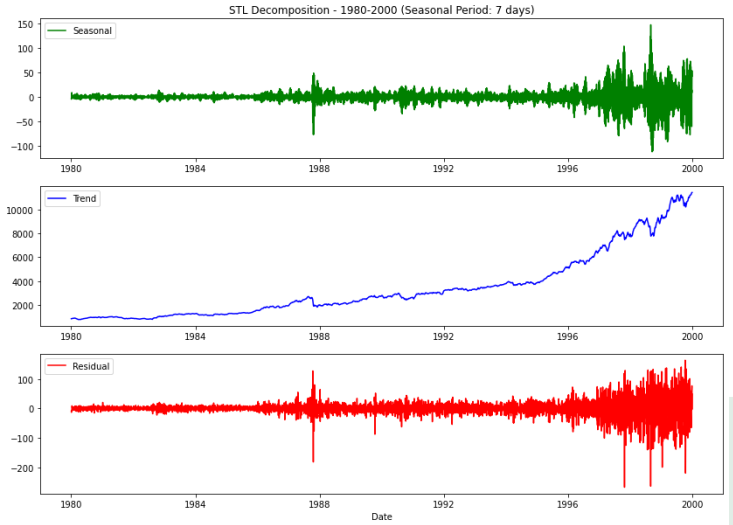

Dow Jones Close Prices Before and After Cleaning (2000 onwards)
Original Close Prices
Cleaned Close Prices
Close Price
Date
5000
10000
15000
20000
25000
30000
35000
40000
2000
2004
2008
2012
2016
2020
2024

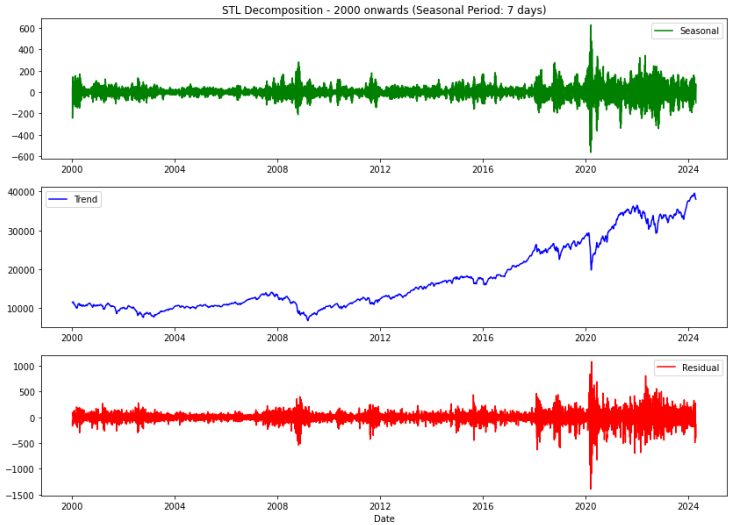
STL Decomposition:

Period: 7 days

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Weekly patterns with increased fluctuations during economic events. • Significant long-term growth with dips during crises. • High volatility during downturns.	• Consistent weekly patterns with spikes during crashes. • Strong upward growth trend. • Increased volatility during market corrections.	• Regular weekly patterns, disrupted during crises. • Significant upward growth. • High volatility during recent crises.



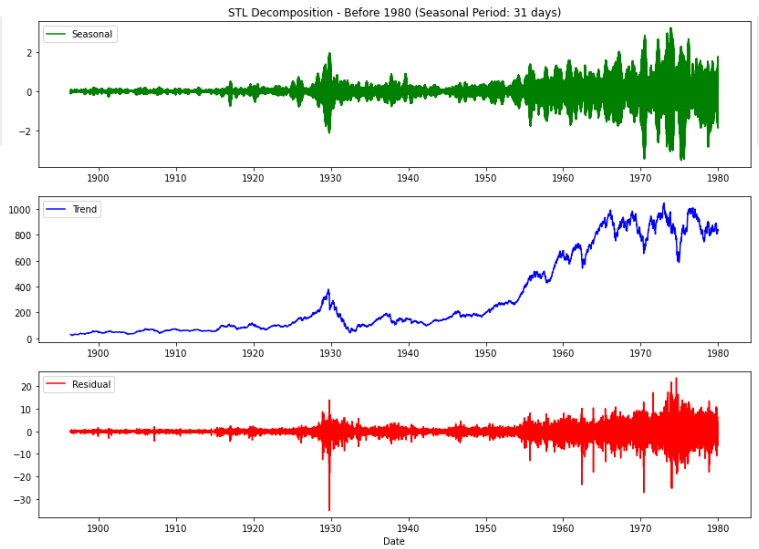


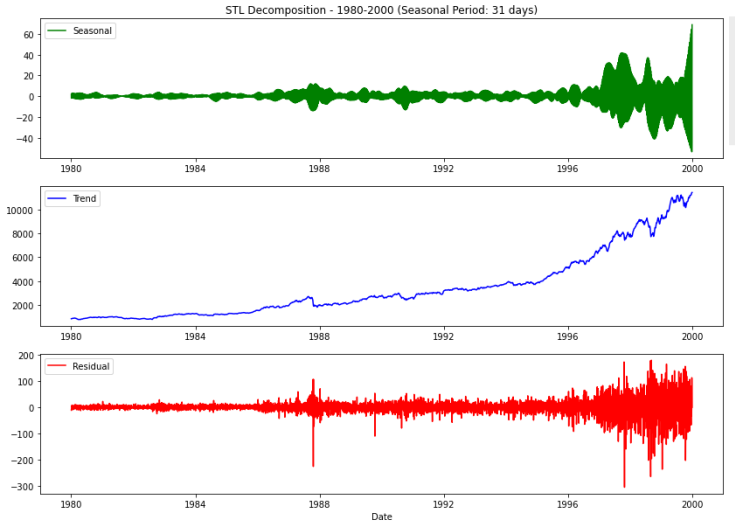


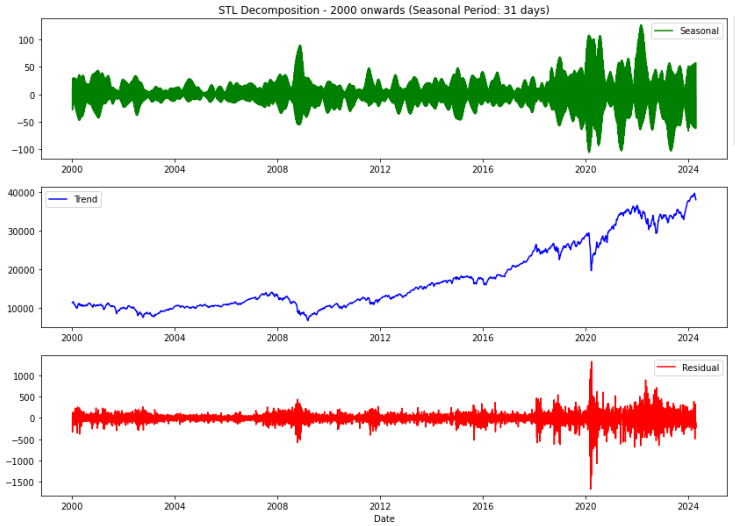
STL Decomposition:

Period: 31 days

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Monthly patterns reflecting economic cycles. • Steady growth with dips during crises. • Higher volatility during downturns.	• Consistent monthly patterns with event-related spikes. • Strong growth trend. • Increased volatility during corrections.	• Regular monthly patterns disrupted during crises. • Significant upward growth. • High volatility during economic events.



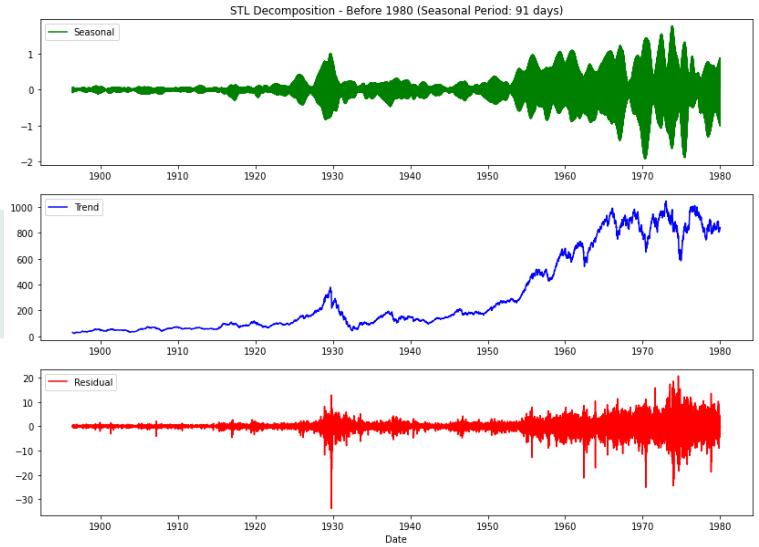


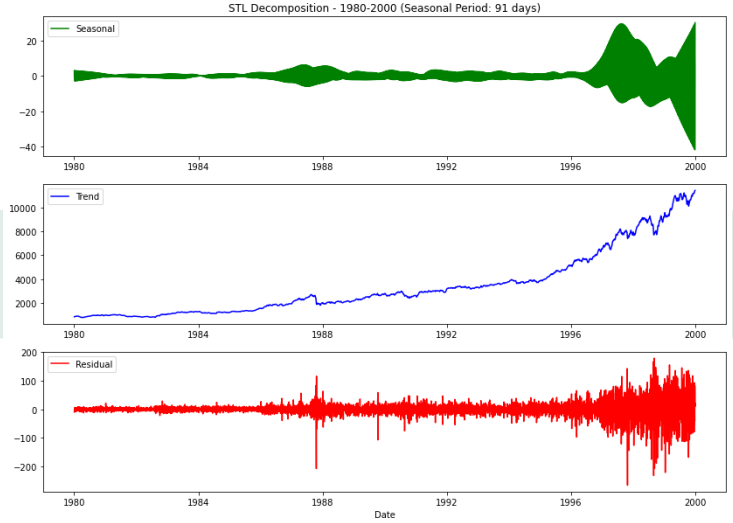


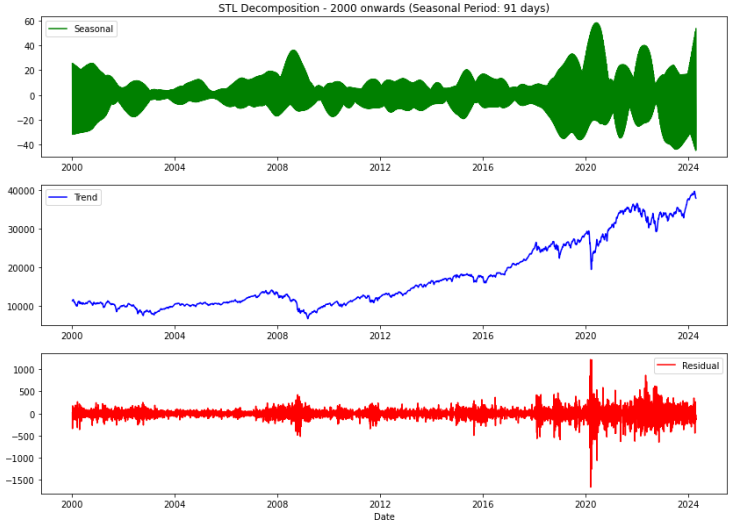
STL Decomposition:

Period: 91 days

Before 1980 Segment:	1980-2000 Segment	2000 onwards Segment
• Quarterly patterns reflecting longer economic cycles. • Steady growth with significant downturns. • High volatility during economic downturns.	• Consistent quarterly patterns with event-related changes. • Strong upward trend. • Increased volatility during corrections.	• Regular quarterly patterns with significant disruptions during crises. • Significant upward growth. • High volatility during global events.







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:

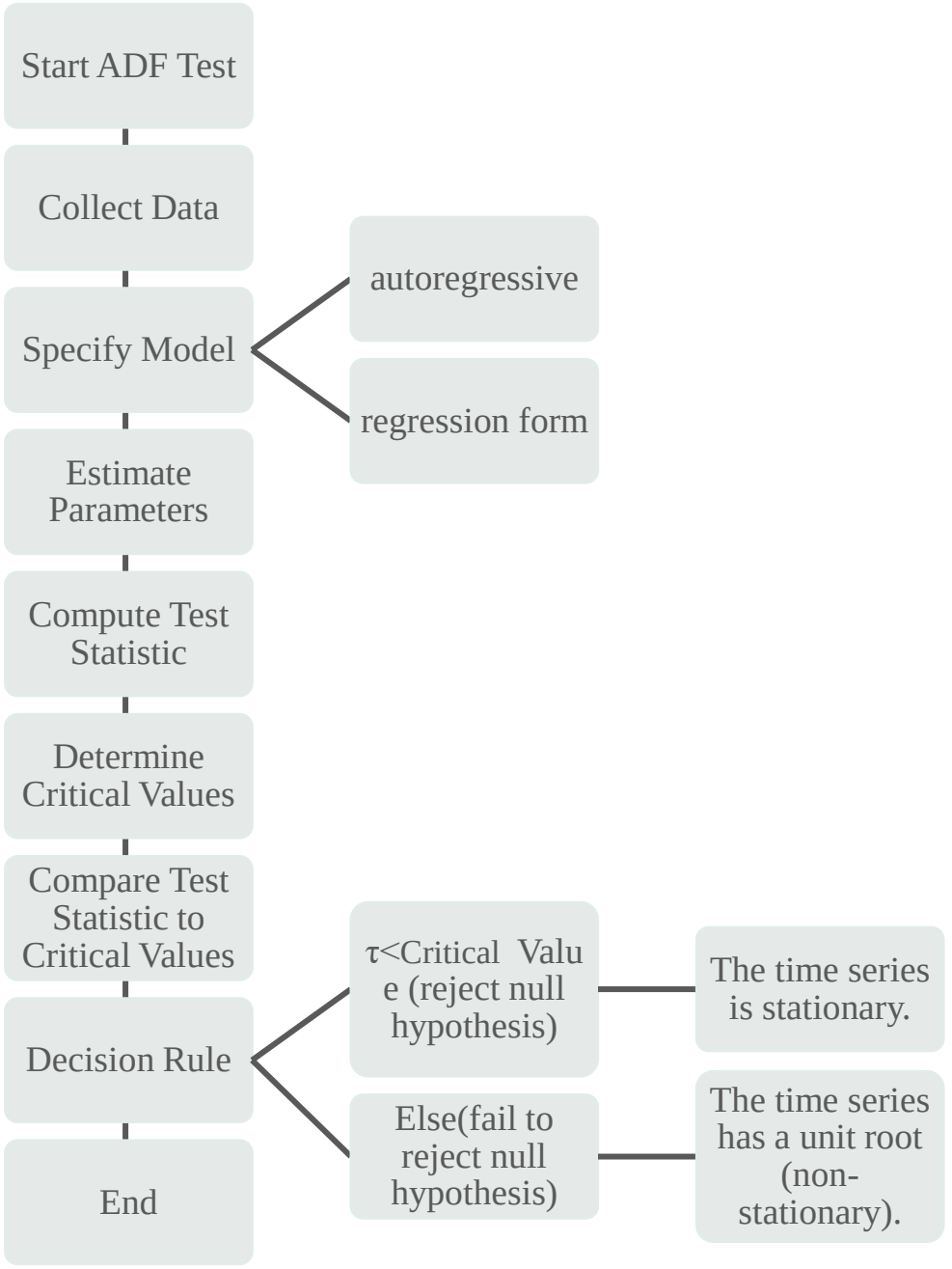
trends, cycles, random walks, or other structures that change over time.

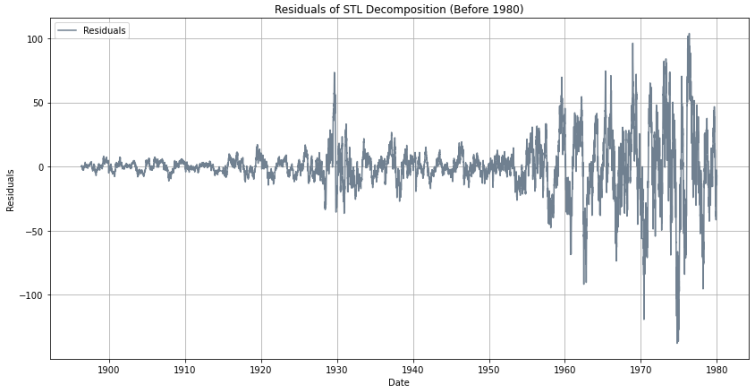
The Dickey-Fuller Test:

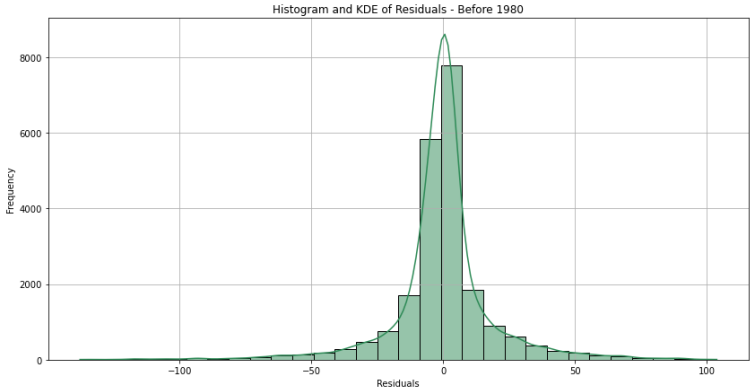
- The Dickey-Fuller test is used to determine whether a time series is stationary or not.
- This test is capable of revealing the presence of unit roots in the series, indicating non-stationarity.

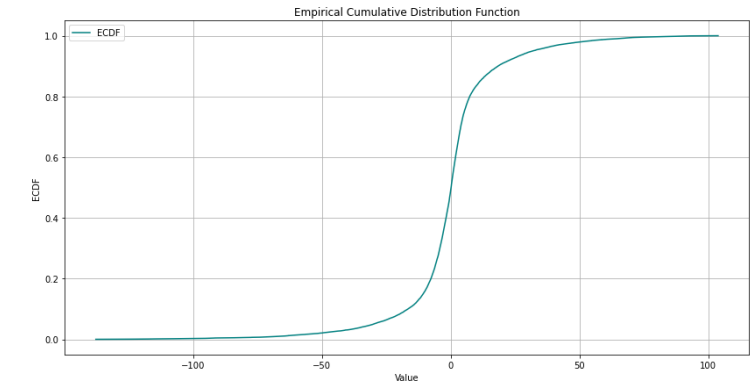
The Null and Alternative Hypotheses:

- 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).









STL Decomposition - Residuals Analysis (Segment: Before 1980):

Residuals of STL Decomposition:

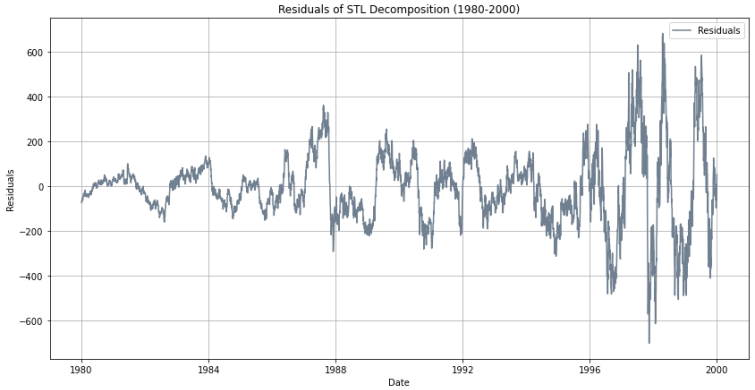
- Observing the residuals plot, we see the variance and volatility before 1980, especially during economic events like the Great Depression.
- The histogram and KDE plot show a concentration of residuals around the mean with some outliers, indicating non-normal distribution.
- The ECDF plot confirms the distribution and extreme values present in the residuals.

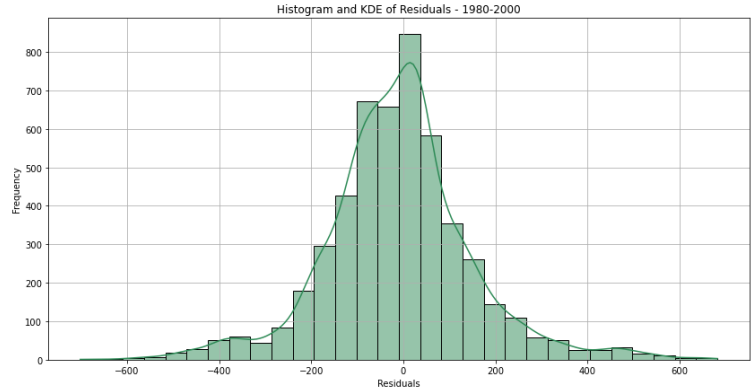
ADF Test:

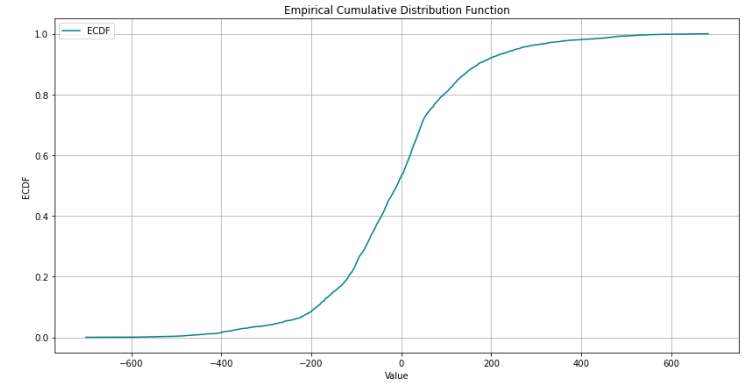
- ADF Statistic: -11.6194
- p-value: 2.4045e-21
- Critical Values: 1%: -3.4306, 5%: -2.8617, 10%: -2.5668
- Conclusion: The residuals are stationary as the p-value is less than 0.05.

Statistics for Residuals:

- Mean: -0.0296
- Median: 0.0835
- Std Dev: 19.8883
- Skewness: -0.4060
- Kurtosis: 7.3487







STL Decomposition - Residuals Analysis (Segment: 1980-2000):

Residuals of STL Decomposition:

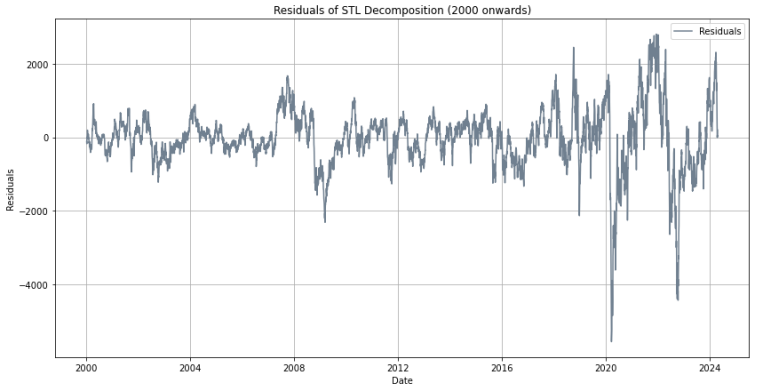
- The residuals plot for 1980-2000 shows increased volatility during market crashes like Black Monday in 1987.
- The histogram and KDE plot display a more spread distribution with outliers, indicating periods of market corrections.
- The ECDF plot confirms the distribution and extreme values present in the residuals.

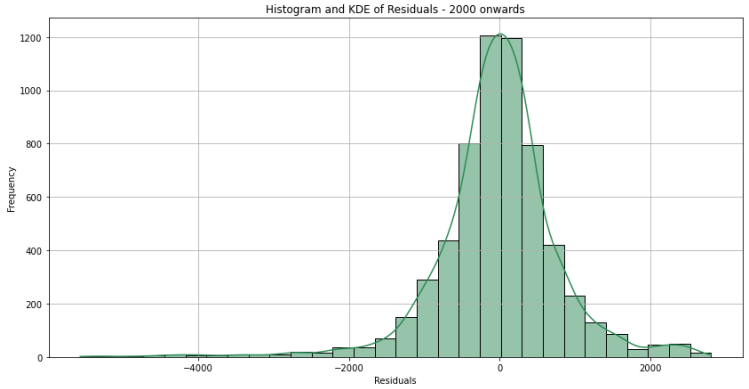
ADF Test:

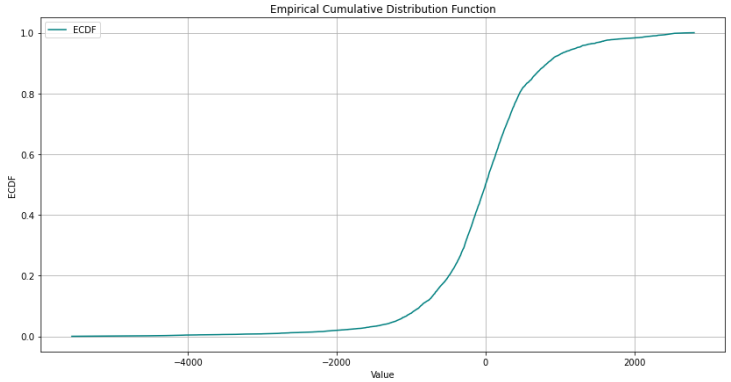
- ADF Statistic: -5.7879
- p-value: 4.9496e-07
- Critical Values: 1%: -3.4317, 5%: -2.8621, 10%: -2.5671
- Conclusion: The residuals are stationary as the p-value is less than 0.05.

Statistics for Residuals:

- Mean: -10.3015
- Median: -9.3663
- Std Dev: 160.6661
- Skewness: 0.2394
- Kurtosis: 1.9452







STL Decomposition - Residuals Analysis (Segment: 2000 onwards):

Residuals of STL Decomposition:

- The residuals plot for 2000 onwards shows extreme volatility during events such as the 2008 financial crisis and the COVID-19 pandemic.
- The histogram and KDE plot reveal a highly skewed distribution with significant outliers, indicating the market's global economic events.
- The ECDF plot confirms the distribution and extreme values present in the residuals.

ADF Test:

- ADF Statistic: -6.4237
- p-value: 1.7661e-08
- Critical Values: 1%: -3.4314, 5%: -2.8620, 10%: -2.5670
- Conclusion: The residuals are stationary as the p-value is less than 0.05.

Statistics for Residuals:

- Mean: -22.6995
- Median: 0.9413
- Std Dev: 818.1939
- Skewness: -0.8748
- Kurtosis: 5.8504

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 specific distribution, typically the 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:

- If the p-value is less than the significance level (e.g., 0.05), reject H_0, indicating that the sample does not follow the reference distribution.

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

- $\mathcal{D}$ 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.

For the period "Before 1980":

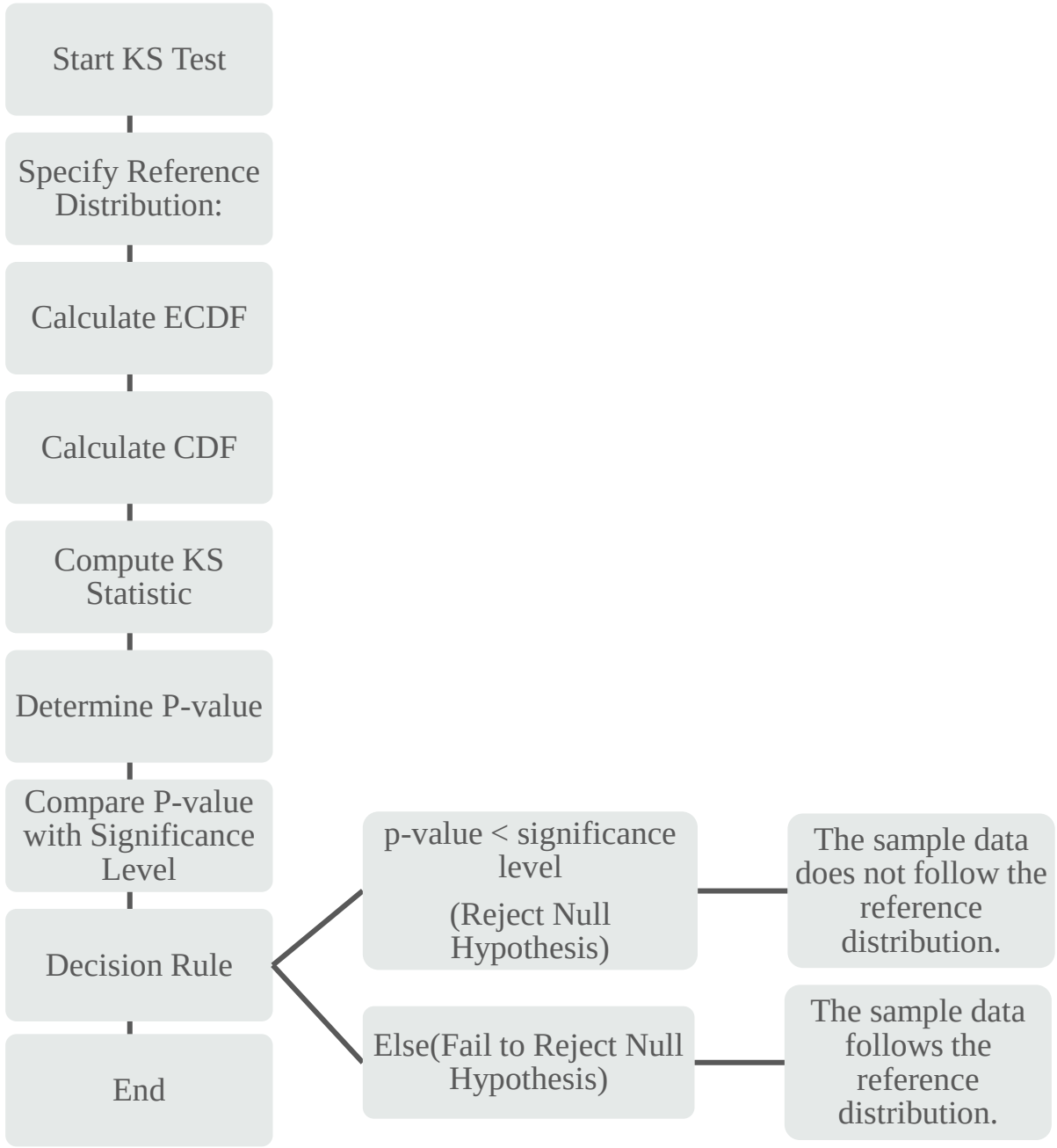
- Statistic: 0.1569
- P-value: 0.0
- Interpretation: The residuals are not normally distributed.

For the period "1980-2000":

- Statistic: 0.0751
- P-value: 3.009792681805 37e-25
- Interpretation: The residuals are not normally distributed.

For the period "2000 onwards":

- Statistic: 0.0868
- P-value: 1.4281487679606139e-40
- Interpretation: The residuals are not normally distributed.



Quantile Normalization of Residuals:

Objective:
To transform residuals so that they follow a standard normal distribution.

Start Quantilization → STL Decomposition → Order Residuals from Smallest to Largest → Compute Quantiles of Standard Normal Distribution → Map Each Residual to Corresponding Quantile → Validate Normality: K-S Test, Q-Q plot, Histogram → End

Mathematical Representation:

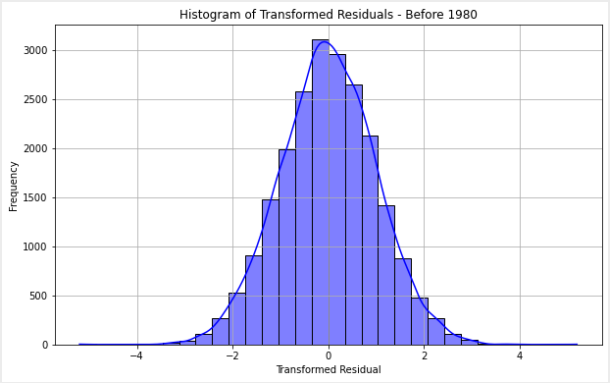
Normalized Residual $= \boldsymbol{\Phi^{-1}}\left(\frac{\boldsymbol{Rank(RESIDUAL)}}{\boldsymbol{n+1}}\right)$

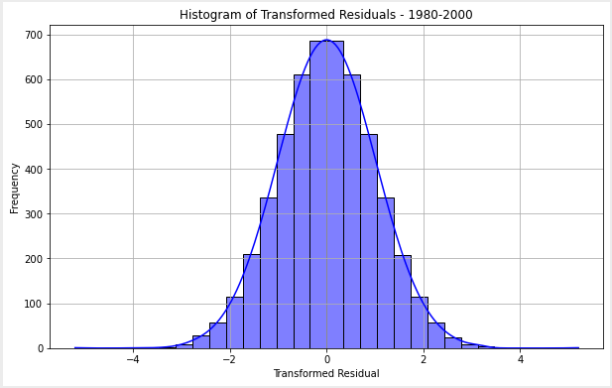
- $\boldsymbol{\Phi^{-1}}$ is the inverse cumulative distribution function (CDF) of the standard normal distribution.
- $\boldsymbol{Rank(RESIDUAL)}$ is the rank of the residual in the ordered list.
- $\boldsymbol{n}$ is the total number of residuals.

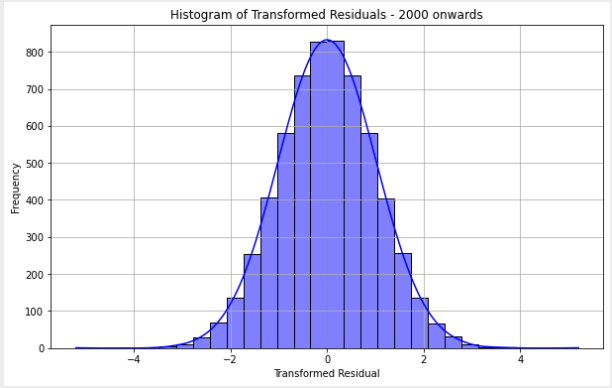
Advantages:

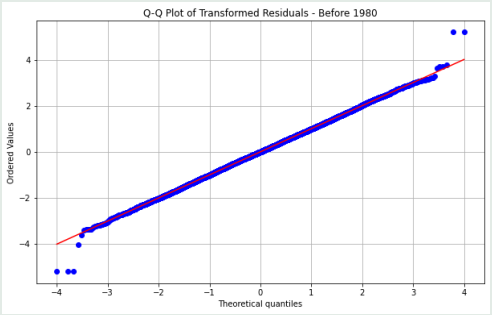
- Reduces the effect of outliers.
- Ensures the residuals follow a normal distribution

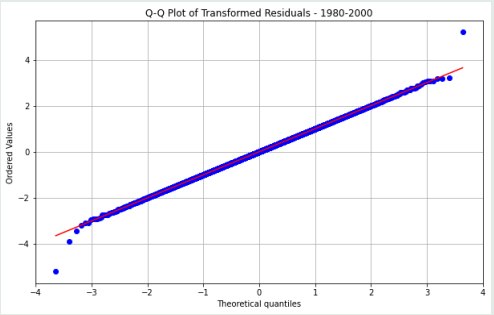
Analysis and Results of Residuals Quantile Normalization:

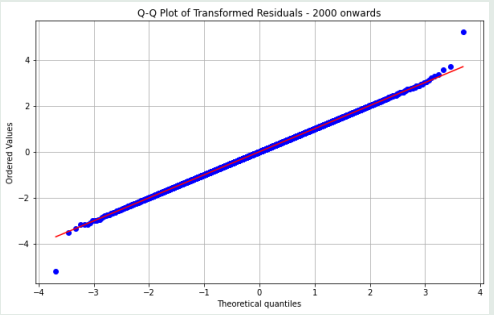












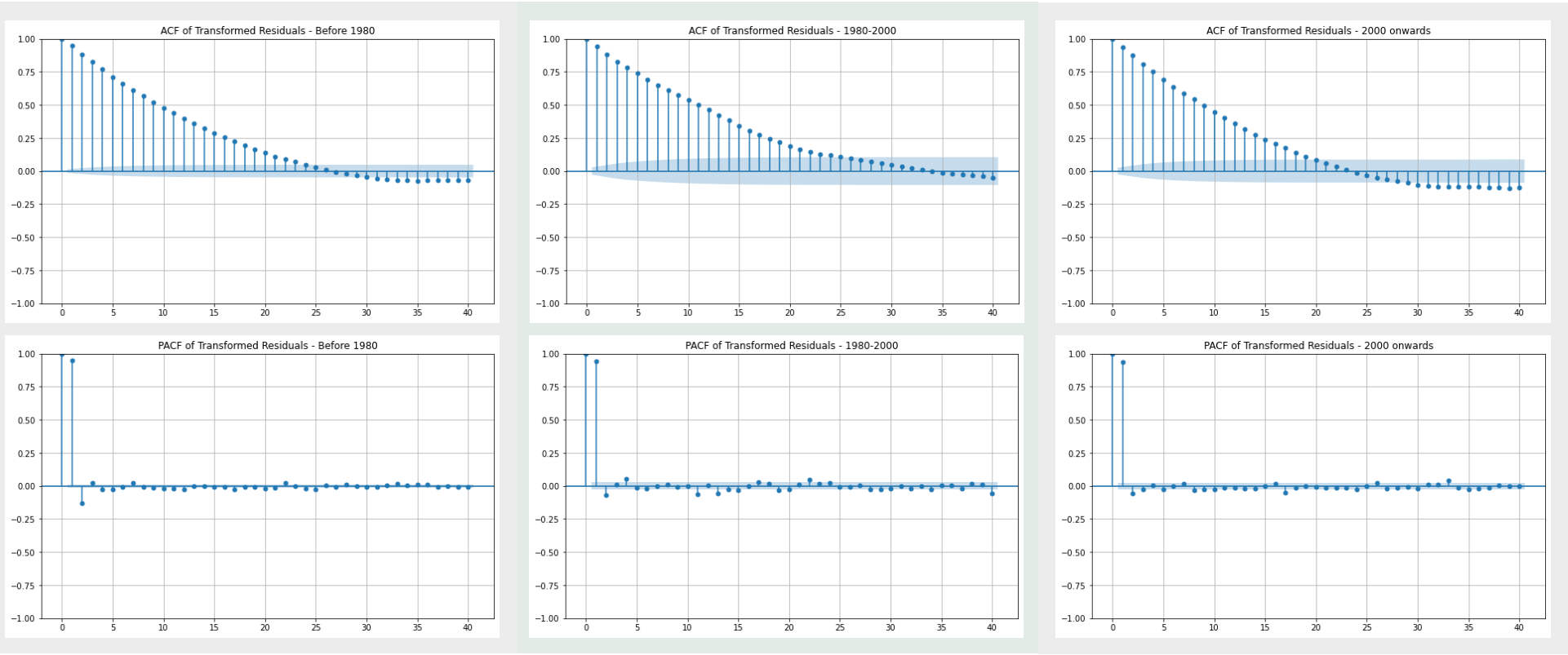
Kolmogorov-Smirnov Test:

- Statistic: 0.00731
- P-value: 0.1893201725285546
- Conclusion: the transformed residuals are normally distributed.

- Statistic: 0.0014520
- P-value: 1.0
- Conclusion: the transformed residuals are normally distributed.

- Statistic: 0.001286
- P-value: 0.8
- Conclusion: the transformed residuals are normally distributed.

ACF and PACF of Transformed Residuals: Before 1980, 1980-2000, and 2000 Onwards:



Result:

Autocorrelation Function (ACF):

- Gradual decline in correlation over lags, indicating autocorrelation across all segments.

Partial Autocorrelation Function (PACF):

- Significant spike at lag 1, with minimal subsequent correlations across all segments.

Introducing Models:

ARMA(p, q) Model

Definition:

- ARMA stands for AutoRegressive Moving Average model.
- Combines two parts: AutoRegressive (AR) part and Moving Average (MA) part.

Formula:

$$Z_t = \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

- Z_t is the time series value at time t.
- ϕ_i are the coefficients of the AR part.
- θ_i are the coefficients of the MA part.
- a_t t is white noise error term.

ARMA Model Identification, Estimation, and Evaluation

CF and PACF:

- ACF (Autocorrelation Function):

Measures the correlation between observations at different lags.

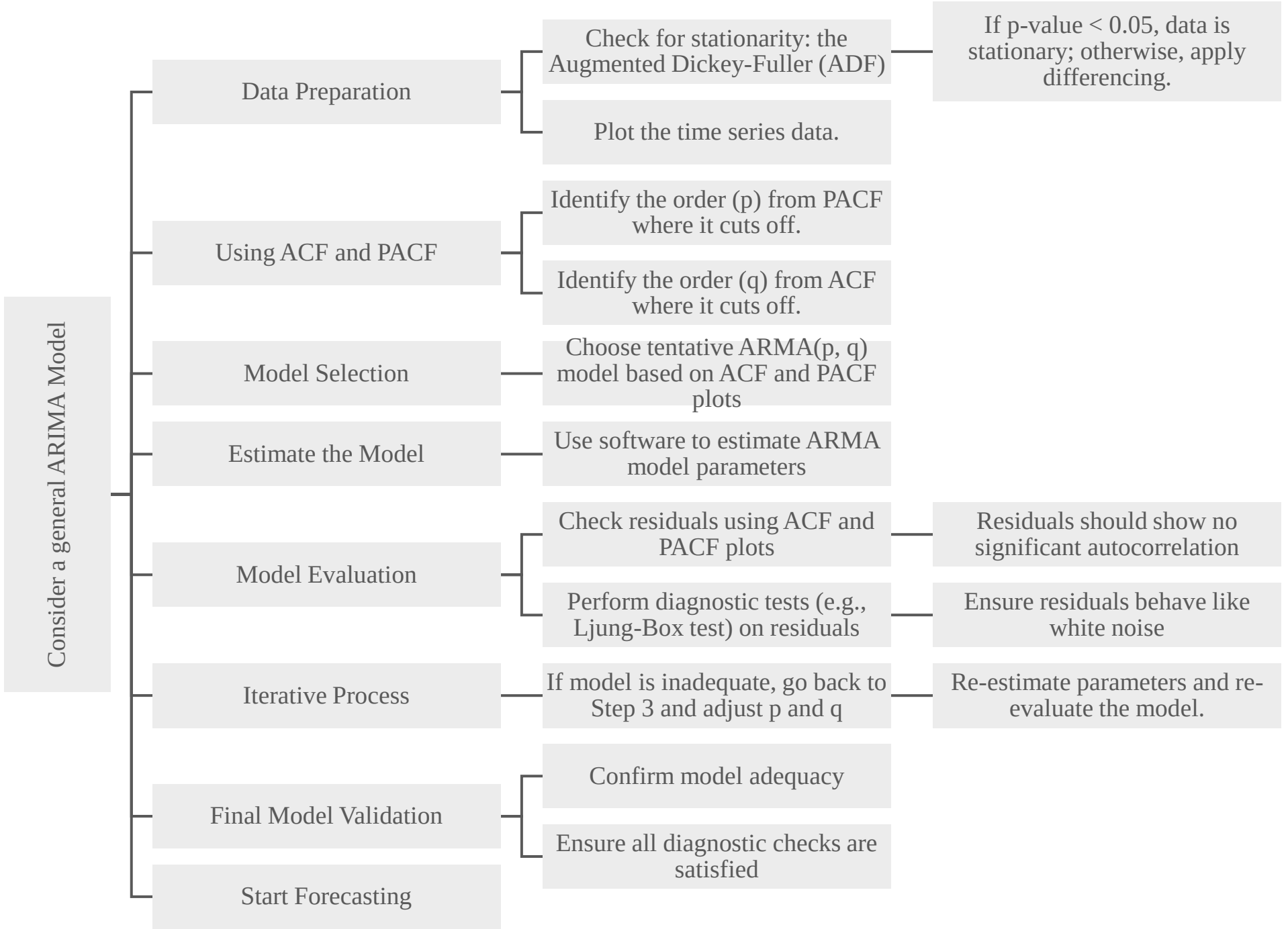
- PACF (Partial Autocorrelation Function):

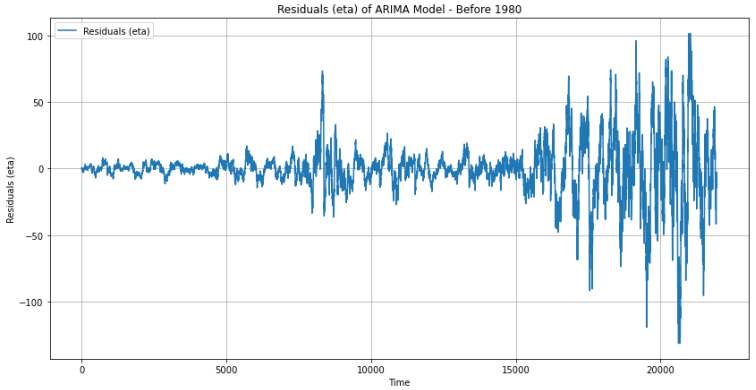
Measures the correlation between observations at different lags after removing the effects of intermediate lags.

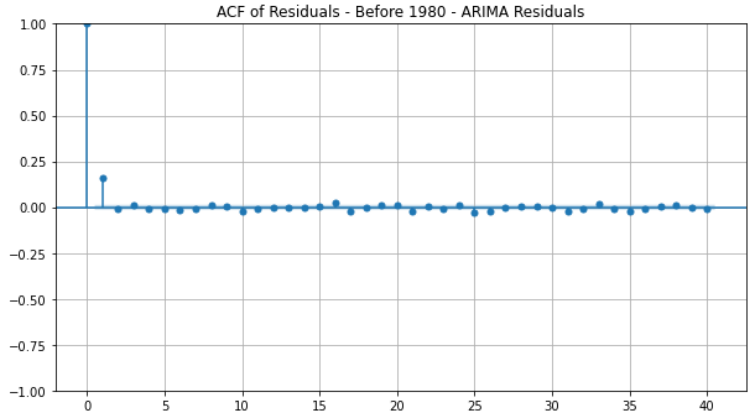
Model Identification Using ACF and PACF:

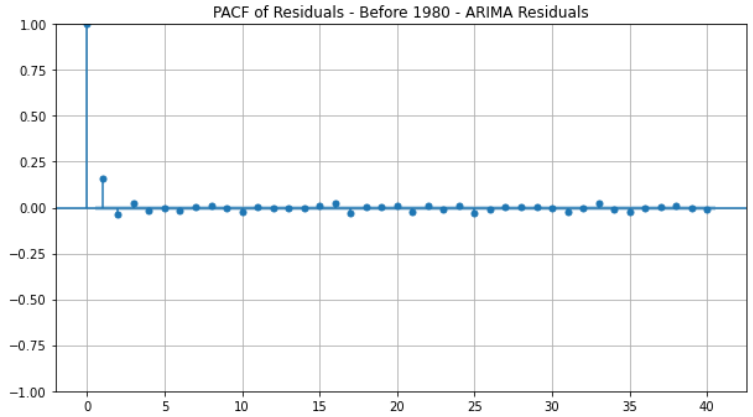
- AR(p): ACF tails off, PACF cuts off after p lags.
- MA(q): ACF cuts off after q lags, PACF tails off.
- ARMA(p, q): Both ACF and PACF tail off.

ARMA Model Flowchart:









ARIMA Model Residuals - Before 1980:

ACF of Residuals:

- The lack of significant autocorrelations beyond lag 1 suggests that the residuals do not exhibit patterns of correlation over time, indicating that the model has successfully captured the underlying data structure.

PACF of Residuals:

- The significant spike at lag 1 with minimal correlations at higher lags further supports that the model has accounted for the primary autocorrelation structure in the data.

Residuals (eta) Plot:

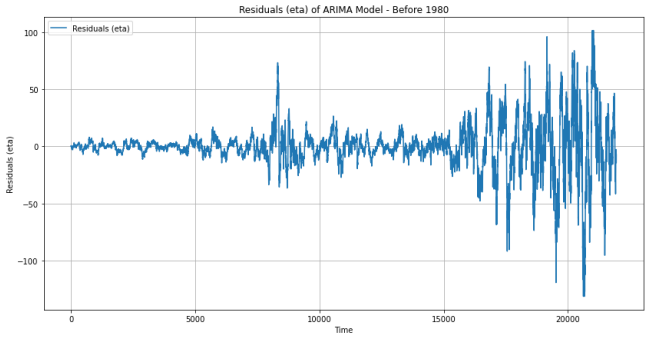
- The residuals behaving like white noise, fluctuating around zero without any discernible pattern, is a strong indication that the residuals are random. This randomness implies that the model has effectively captured the information in the data, leaving only random noise as residuals.

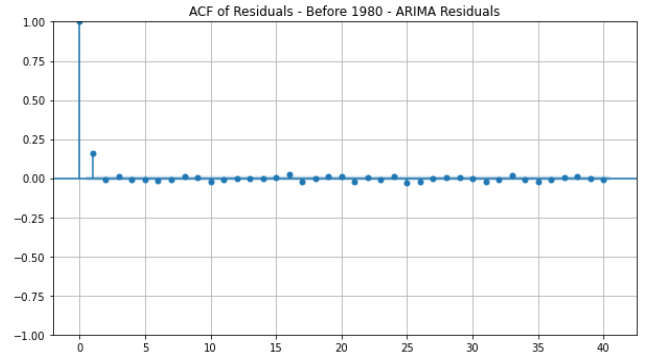
Model Diagnostics:

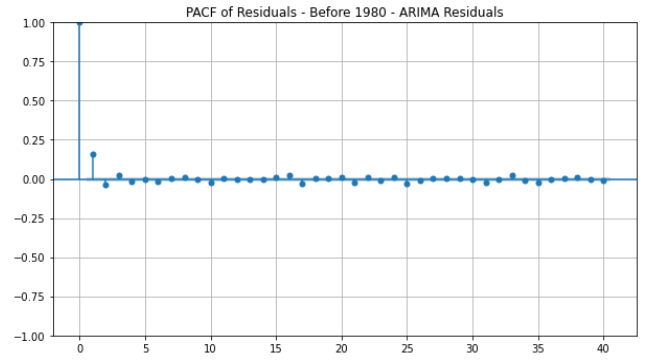
- Minimal autocorrelation in the residuals is a key diagnostic check for a good ARIMA model fit.The residuals centered around zero with no obvious pattern suggest that there are no systematic errors left unmodeled.

ARIMA Model Residuals:

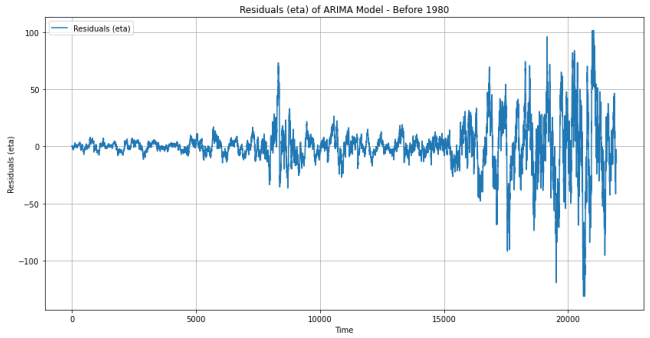
Before 1980:

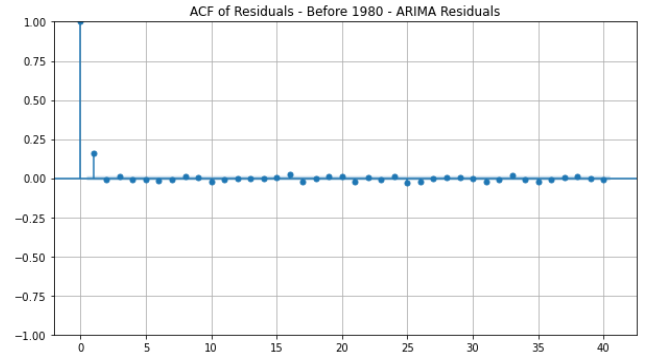


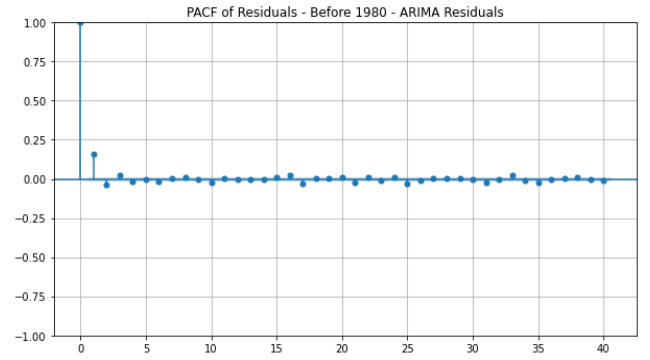




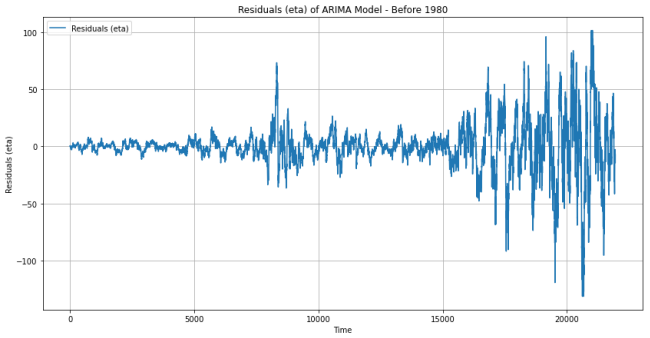
1980-2000:

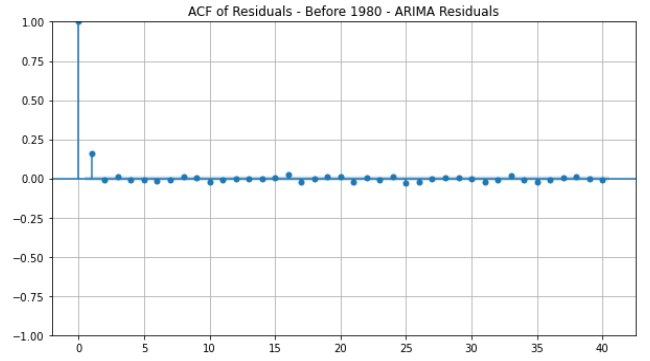


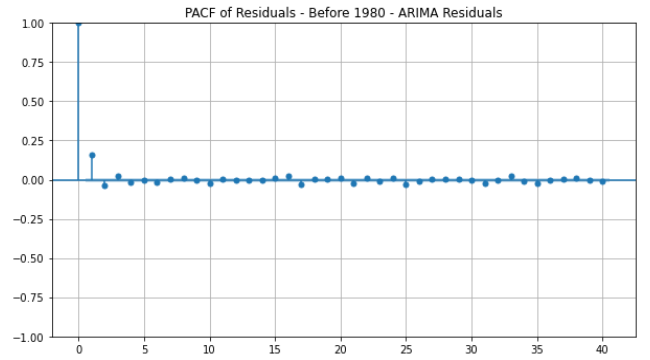




2000 Onwards:







ACF and PACF of Residuals:

- No significant autocorrelations beyond lag 1 in ACF; significant spike at lag 1 in PACF indicates effective model capture.

Residuals (eta) Plot:

- Residuals behave like white noise, implying the model has captured the data's information, leaving only random noise.

Model Evaluation and Validation: 1980-2000:

Model Evaluation:

- The ACF of residuals shows no significant autocorrelations beyond lag 1.
- The PACF of residuals has a significant spike at lag 1, indicating minimal correlation at higher lags.
- Residuals exhibit minimal autocorrelation, suggesting a good model fit.

Model Validation:

The Ljung-Box Q-statistics for all segments showed p-values indicating no significant autocorrelation in the residuals:

•Before 1980: Q-statistic = 553.74, p-value = 0.00.

•1980-2000: Q-statistic = 3.18, p-value = 0.07.

•2000 Onwards: Q-statistic = 2.29, p-value = 0.13.

✓ **Overall, the residuals meet the criteria for minimal autocorrelation, validating the model's adequacy.**

Conclusion:

Model Adequacy:

- The ARIMA(1,0,0) model effectively captured the underlying data structure, with residuals showing minimal autocorrelation across all time segments.

Diagnostic Tests:

- Ljung-Box Q-statistics confirmed that residuals are uncorrelated and exhibit constant variance, validating the model's adequacy.

Predictive Reliability:

- Given the minimal autocorrelation in residuals and successful diagnostic tests, the ARIMA(1,0,0) model is suitable for reliable forecasting of future stock prices.

Future Work Suggestions - Forecasting

Scenario Analysis:
- Assess impact of various economic conditions on stock prices.

Prediction Intervals:
- Provide a range of possible future values to capture uncertainty.

Advanced Models:
- Explore deep learning models like LSTM for non-linear dependencies.
- Combine ARIMA with machine learning techniques.

Real-Time Forecasting:
- Implement automated pipelines for data collection and forecasting.
- Develop interactive dashboards for real-time updates.

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