Exploratory Data Analysis Approach to Identify and Analyze a Systemic Financial Risk

Data Exploration and Visualization on S&P500

Win Htet

10 - MARCH - 2024

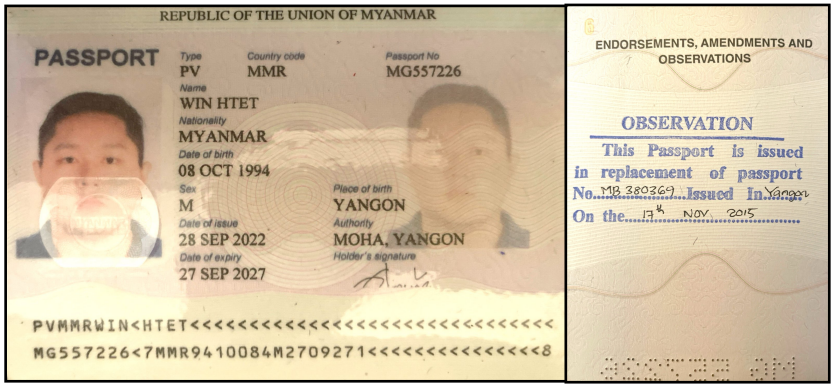
Group #4885

Group Members

- *Win Htet*
- *Hnin Oo Wai*

Government Issued Photo ID

- Name : Win Htet (Mr.)
- NRC No.: 10/PAMANA(N)178645
- Passport : MG557226 (former MB380369)



Renewed Passport

Myanmar Driving License

*** NRC - National Registration Card*

Project Goal and Importance

Project Goal:

- Utilize EDA to identify and analyze systemic risk vulnerabilities in global financial markets, focusing on the U.S

Project Goal and Importance

Importance of the Project:

- Identify / Analyze of systemic risk with POV from third-world countries
- Create data visualization by applying Quantitative approach
- Drive future research in risk analysis

Literature Review/Background

Summarize Table for past Literature Review:

Table 1. Past Literature Reviews Summary

Sr. No.	Author(s) (Year)	Problem Type	Model Applied	Dataset Used
1	Ma et al., (2022) [1]	Identifying systemic financial risks before and after the COVID-19 outbreak	Copula-GARCH with CES approach	Financial data from the US market
2	Sun et al., (2009) [2]	Identifying systemic events at an early stages and containing steps	Financial soundness indicators, advanced tools like CCA, option iPoD, Multivariate Dependency, etc	Financial data from the US market
3	Craig, (2020) [3]	Developing a systemic risk indicator	Not specified	US financial market data
4	Hullman & Gelman, (2021) [4]	Exploring systemic risks indicator	Exploratory Data Analysis (EDA)	Various financial datasets
5	Adamjee, (2023) [5]	Analyzing stock market data for systemic risks	EDA techniques, statistical analysis	Historical stock market data

Literature Review/Background

Relation to Existing Work:

- Especially, [Hullman & Gelman, 2021] and [Adamjee, 2023] make good references for us to continue our capstone project.
- Following their research, we decided to use various financial related datasets and historical dataset of S&P500 to perform EDA with visualization and basic statical approach like ARIMA.

Assumptions and Choice of Technology

Assumptions:

- S&P500 index to represent the US economy situation
- GDP growth rate, federal fund rate, inflation rate and Unemployment rate as external economic indicators that is likely to influence S&P500
- 2002 and 2008 financial crisis possess the characteristics of Systemic risk

Assumptions and Choice of Technology

Table 3. Datasets and Descriptions

Sr. No.	Index	Description	Period	Frequency	Source
1	S&P500	Adjusted Closing Prices of S&P 500	From 1980-01-01 to 2024-01-01	Daily	Yahoo Finance
2	GDP_ Growth_ Rate	Real percent change of GDP growth rate	From 1980 to 2024 (current)	Annual	IMF Datamapper [link]
3	Inflation_ Rate	Inflation rate calculated on average of consumer prices	From 1980 to 2024 (current)	Annual	IMF Datamapper [link]
4	Unemployment_Rate	Unemployment percent of total labor force	From 1980 to 2024 (current)	Annual	IMF Datamapper [link]
5	Interest_rate	Average of Federal Fund Rate	From 1980 to 2024 (current)	Annual	FRED [link]

11,114 data points

45 data points each

Assumptions and Choice of Technology

Choice of Technology:

- Python Programming to perform visualization and analysis
- Required historical data about economic indicators are imported into the program as comma-separated values file.
- ARIMA Diagnostics to understand the events of S&P500 that happen over a period of time
- Univariate, bivariate and multivariate data exploration methods

Assumptions and Choice of Technology

Table 2. Three Basic Data Exploration methods

Sr. No.	Analysis	Variable Relationship	Purpose of method	Techniques
1	Univariate	Single variable	To comprehend the distribution of the variable, detect outliers, and discern any inherent patterns or trends	Histogram, box plot, descriptive statistics (like mean, median, standard deviation)
2	Bivariate	Two variables	To determine how these variables are related and whether a correlation exists	Scatter plots, correlation analysis
3	Multivariate	Three or more variables	To expose deeper understanding, patterns and trends within the dataset	Principal Component Analysis (PCA), cluster analysis

Peer Review Feedback

Our group received the main four points of feedback from peer review.

1. **To apply detailed event-based analysis**
2. **To utilize tables and graphs**
3. **To improve code summary reporting**
4. **To improve writing and formatting style**

Actions Peer Review Feedback

1. To apply detailed event-based analysis

- We added the exact reference year in every graph, chart, plot and included clear explanation in writing format. We also took 2000-2002 dot-com crash and 2008 financial crisis period as the reference behaviors for the system risk and apply them to the 2019 pandemic period.

Actions Peer Review Feedback

2. To utilize tables and graphs

3. To improve code summary reporting

4. To improve writing and formatting style

- We used tables to summarize our main focusing points of the paper and various graphs for visualization.
- We also added detailed explanation on the code and generated results for better understanding.

Summary of Results (Univariate)

Findings:

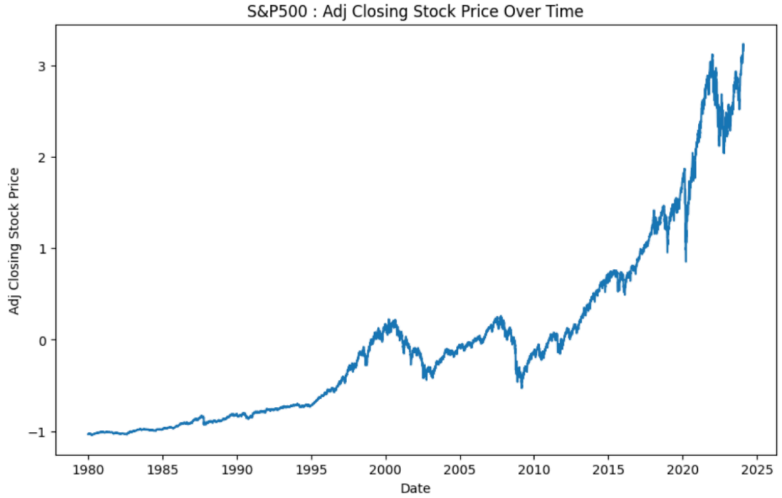


Figure 1. S&P 500 Graph on Adj Closing Price (Daily)

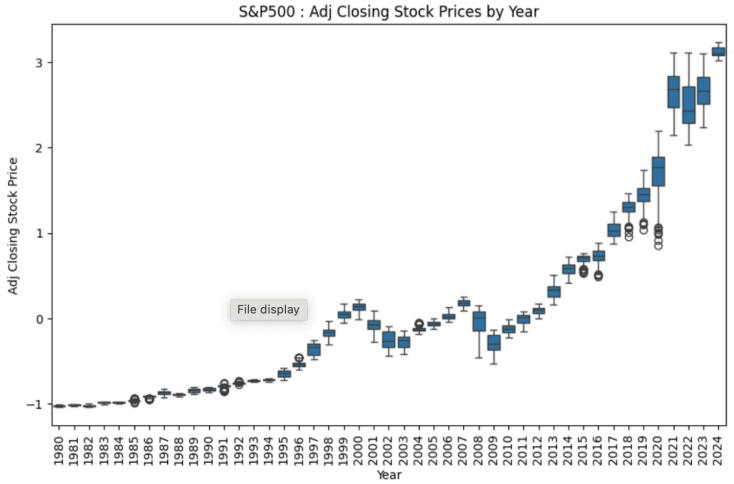


Figure 2. Candle Graph on Adj Closing Price (Daily)

- From 2018 to 2020, the price increased overtime with more outliers in 2019. Recovery time was different for 2008 & 2019.

Summary of Results (Univariate)

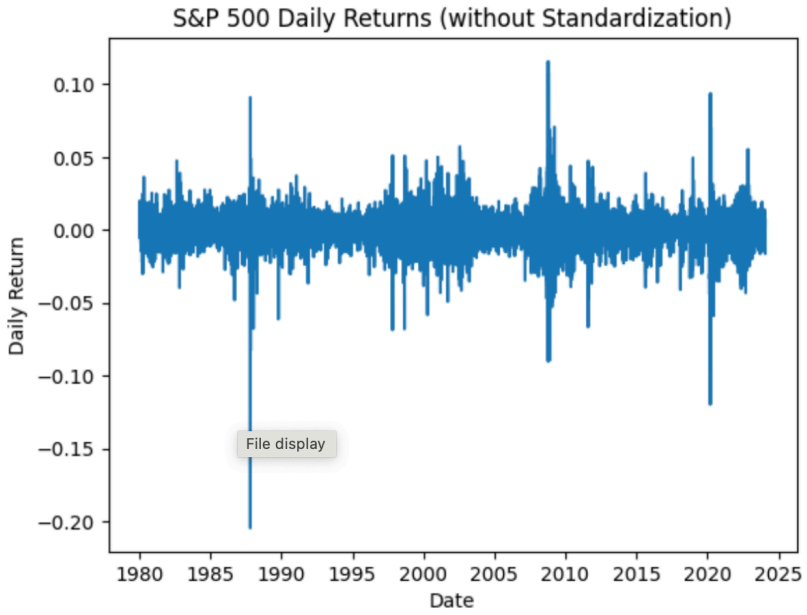


Figure 4. S&P 500 Daily Returns

Findings upon Daily Return:

- Huge spikes in 1987 (stock market crash), 2008 (financial crisis), 2000-2002 (dot-com crash), 2019-2020 (COVID pandemic)
- Great chance for systemic risk, yet, not concluded.

Summary of Results (Bivariate)

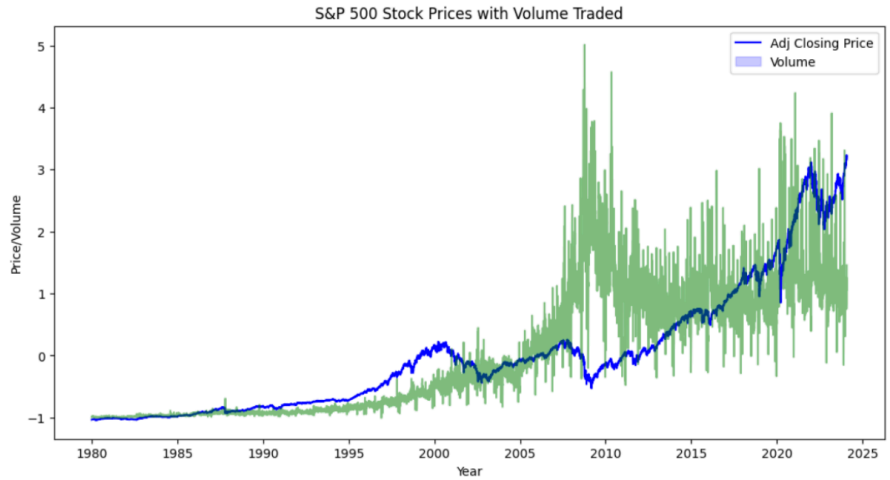


Figure 5. S&P 500 Stock Prices with Volume Traded

Findings upon Bivariate Graph:

- The traded volume spikes with price drop specifically in 2008.
- The price drop and recovery period is quite short in 2019.
- 70.85% positive corr exist in Price vs Traded Volume

Summary of Results (Multivariate)

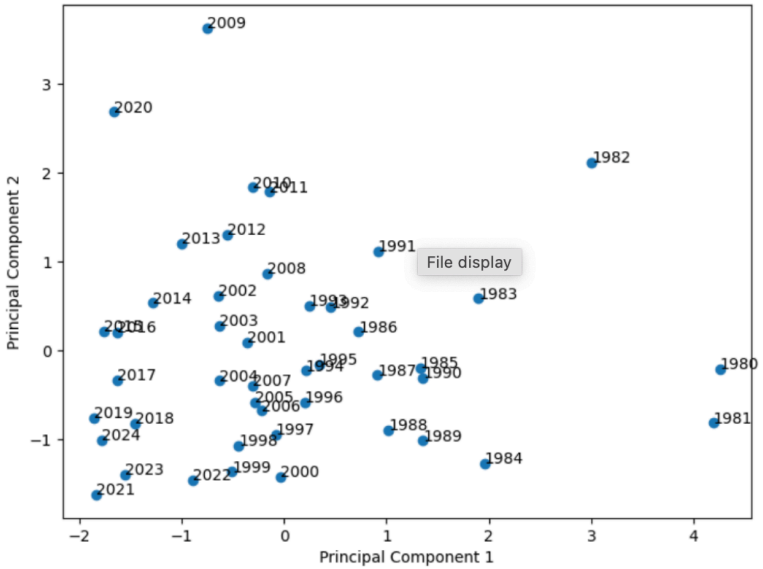
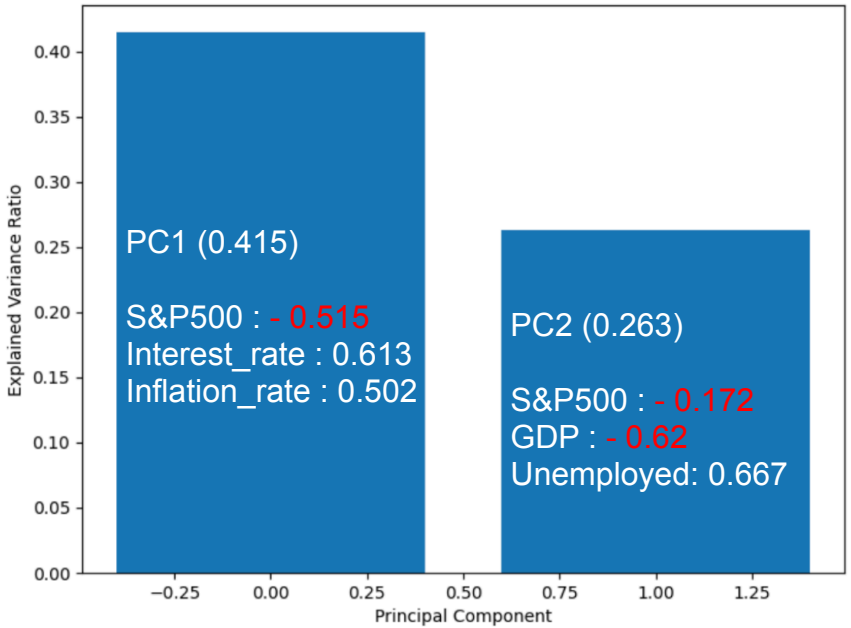


Figure 6. Scatter Plot on the principal components

Findings upon PCA:

- 2002 dot-com crash and 2008 financial crisis seems similar systemic risk profile as they are close to each other.
- COVID-19 pandemic year 2019 locates different position.

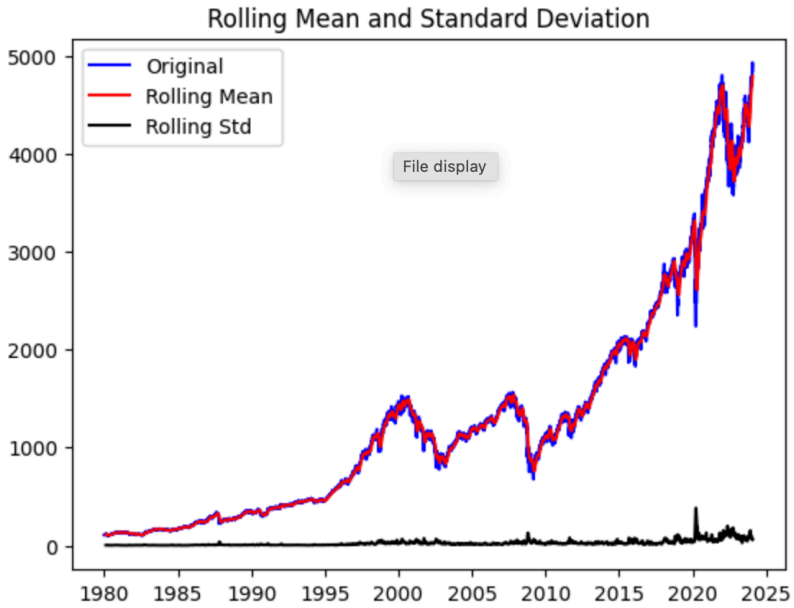
Summary of Results (Multivariate)



Findings upon PCA bar plot:

- 1st principal component (41.5%) is stronger than the 2nd one (26.3%).
- These two components explain the total 67.8% of the total variance.
- Remaining variance is explained by other factors.

Summary of Results (ARIMA)



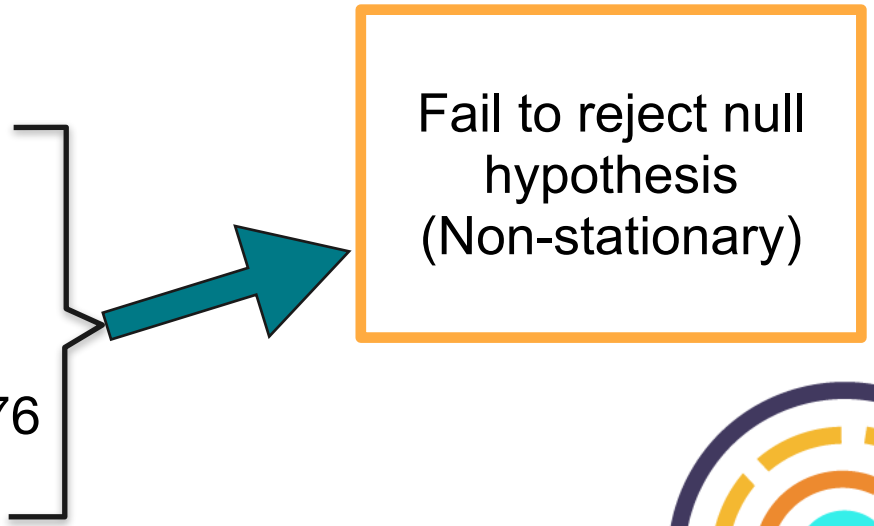
Findings on Moving Average (30):

- Rolling windows of 30 with mean value.
- The mean and standard deviation seem to increase over time.

Summary of Results (ARIMA)

Testing stationary or non-stationary in time-series by the Dickey-Fuller test considering with five facts:

- Test statistics : 2.418
- P-value : 0.999
- Number of lags used : 37
- Number of observation used : 11076
- Critical Values 10% : -2.567



Summary of Results (ARIMA)

Performance of the ARIMA model:

- Log-transform the value as non-stationary
- Train and test data splitting as 90-10

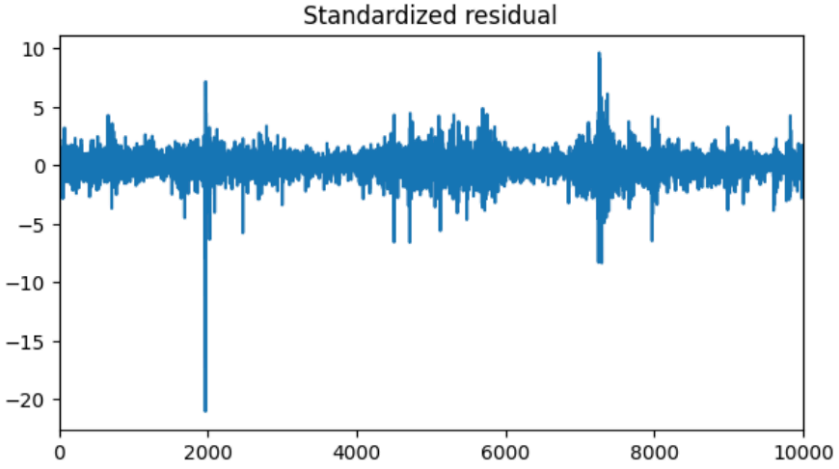
Summary of Results (ARIMA)

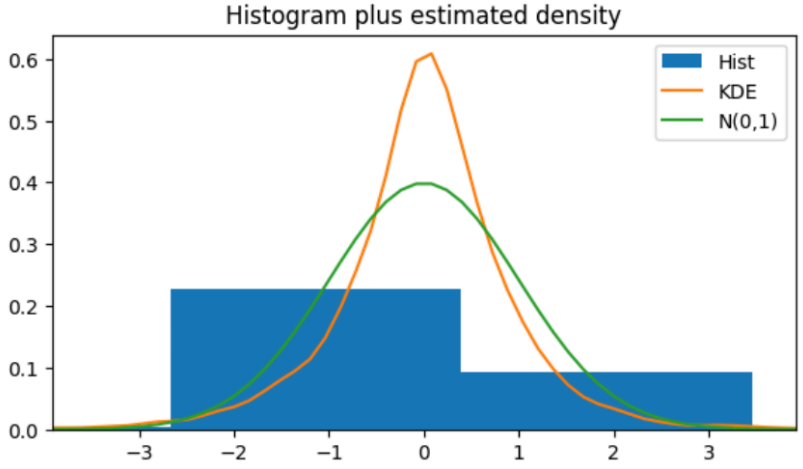
Performance of the ARIMA model:

- Log-transform the value as non-stationary
- Train and test data splitting as 90-10
- After diagnosed, **ARIMA(2,1,0)** is the best model

Summary of Results (ARIMA)

Performance of the ARIMA model (Auto-diagnosis plots):

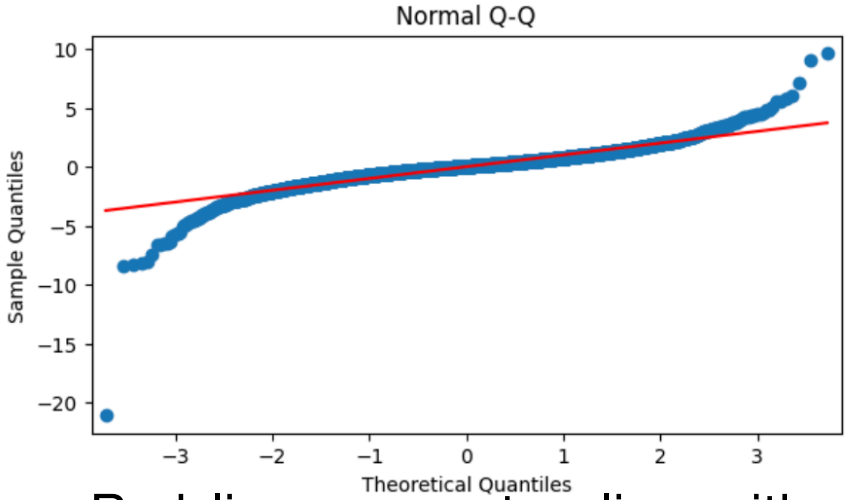


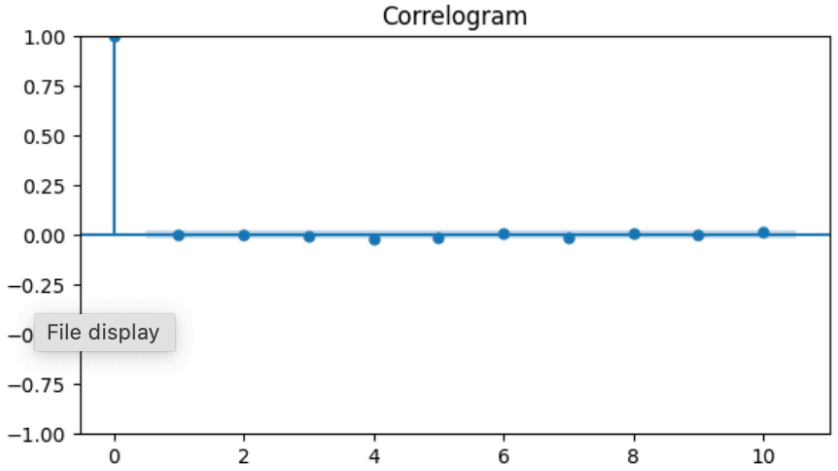


- Standardized Residual seems to possess mean 0 and var 1.
- Density Histogram follows normal distribution.

Summary of Results (ARIMA)

Performance of the ARIMA model (Auto-diagnosis plots):





- Red line seems to align with most of the dots.
- No significant point existed.

Summary of Results (ARIMA)

Forecast Report Performance on Test Data:

- Mean Squared Error (MSE) : 0.1026
- Mean Absolute Error (MAE) : 0.288
- Root Mean Squared Error (RMSE) : 0.32
- Mean Absolute Percentage Error (MAPE) : 0.0346

MAPE of 3.5% suggests that the model is 96.5% accurate in predicting the next sequential observations.

Your Contribution

- Most of the tasks were performed and decided as a group's mutual interest using collaboration platforms like Google CoLab, Google Drive, etc.
- Past Literature Reviews (last three studies)

Your Contribution

- Most of the tasks were performed and decided as a **group**'s mutual interest using collaboration platforms like Google CoLab, Google Drive, etc.
- Past Literature Reviews (last two studies) (individual)

Sr. No.	Author(s) (Year)	Problem Type	Model Applied	Dataset Used
4	Hullman & Gelman, (2021) [4]	Exploring systemic risks indicator	Exploratory Data Analysis (EDA)	Various financial datasets
5	Adamjee, (2023) [5]	Analyzing stock market data for systemic risks	EDA techniques, statistical analysis	Historical stock market data

Your Contribution

- Narrative writing (individual)
  - Abstract
  - 1. Introduction
- Selection of datasets was performed as a **group**
- Program writing on Final Source Code (individual)
- Writing and formatting of Final Source Code (individual)
- Findings and Results conclusion as a **group**

Conclusions

- Multivariate Data Exploration method like PCA can identify and analyze the systemic risk, yet not suitable for predicting exact time and intensity of the risk.
- General time-series analysis approach, such as ARIMA, may not be enough for reliably anticipating trends and unique occurrences in real-world financial circumstances.

Conclusions

- Our project successfully achieved its objectives by leveraging EDA methodologies to identify early warning signs of systemic issues, evaluate financial stability indicators, and analyze market volatility patterns.
- More advanced and complicated approaches such as neural networks and machine learning are more suitable for predicting and managing risks.

Future Scope of Work

- To study about the east Global market focusing on the China market which is emerging in global.
- To study how to manage systemic risk (as now focusing on only identifying and analysis)
- To study more about market volatility by using cross-industry or cross countries analysis
- To accompany neural network and ML for more accurate prediction on future trends and events.

Thank you for your attention!

10 - MARCH - 2024