ĐẠI HỌC KINH TẾ TÀI CHÍNH

FINAL REPORT

Project name: Using seasonal model and missing data estimation to forecast Ceramic Store Revenue in Time series analysis

Group: 9
Class: A01E
Instructor: MsC. Ngo Thuan Du

Ho Chi Minh city, Thursday 30th October 2025

Group members

1	Phùng Nguyễn Ngọc Minh	225210806	100
2	Nguyễn Hoàng Vân Khánh	225210732	100
3	Phạm Nguyễn Uyển Nhi	225210784	100
4	Đinh Thị Thu Trang	225210681	100
5	Nguyễn Thúy Vy	225210789	100

Table of contents

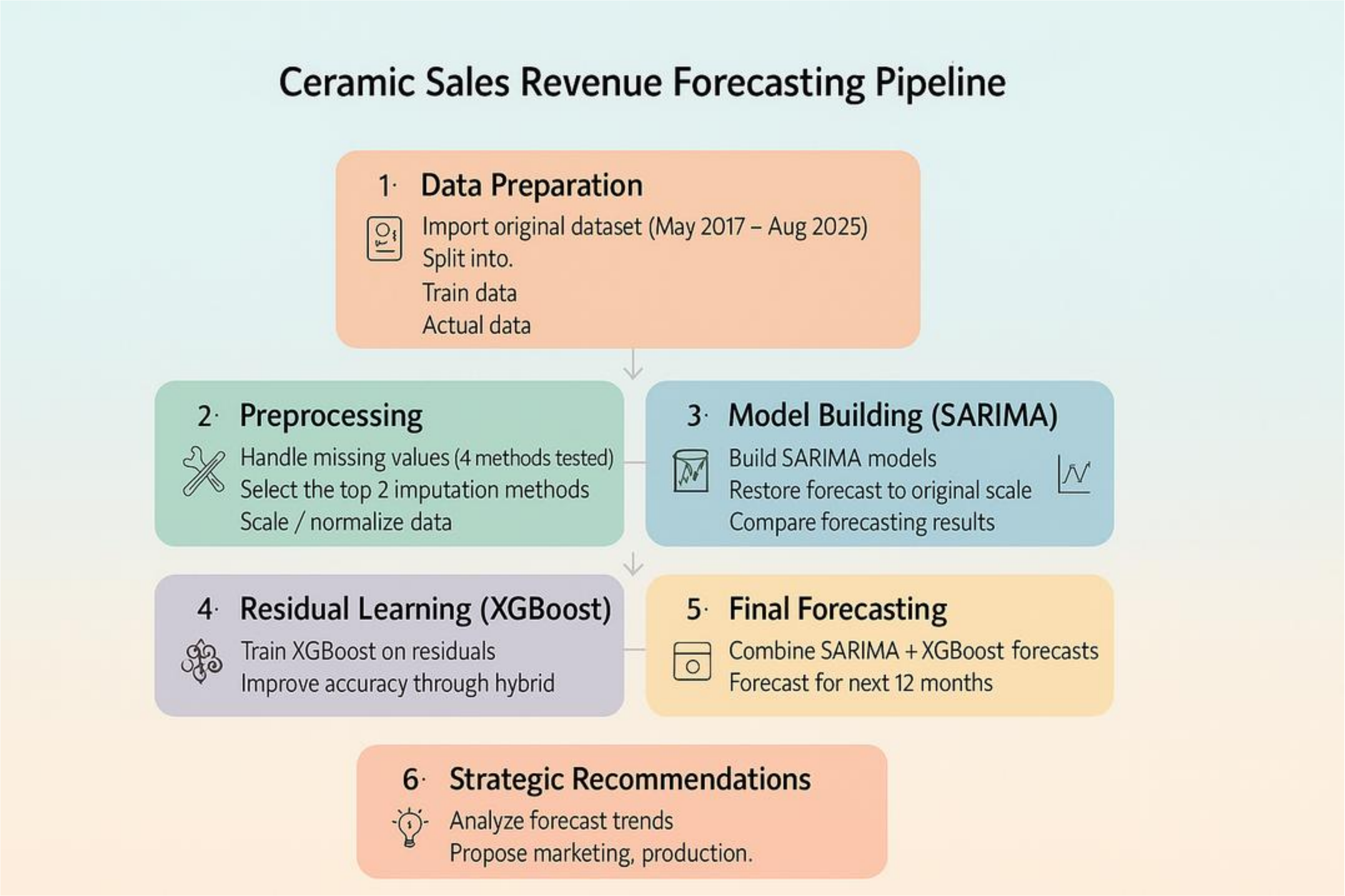
I. INTRODUCTION

1. **Reason for Topic Selection:** The ceramics industry is characterized by strong seasonality and significant fluctuations throughout the year. This necessitates that businesses accurately forecast revenue to proactively manage production and business operations.
2. **Research Objectives:**

- Develop and evaluate various ceramic revenue forecasting models to identify the model with the highest accuracy.
- Apply the forecasting results to support businesses in planning production, distribution, and business strategies according to seasonal cycles.

1. **Tool Used:** R

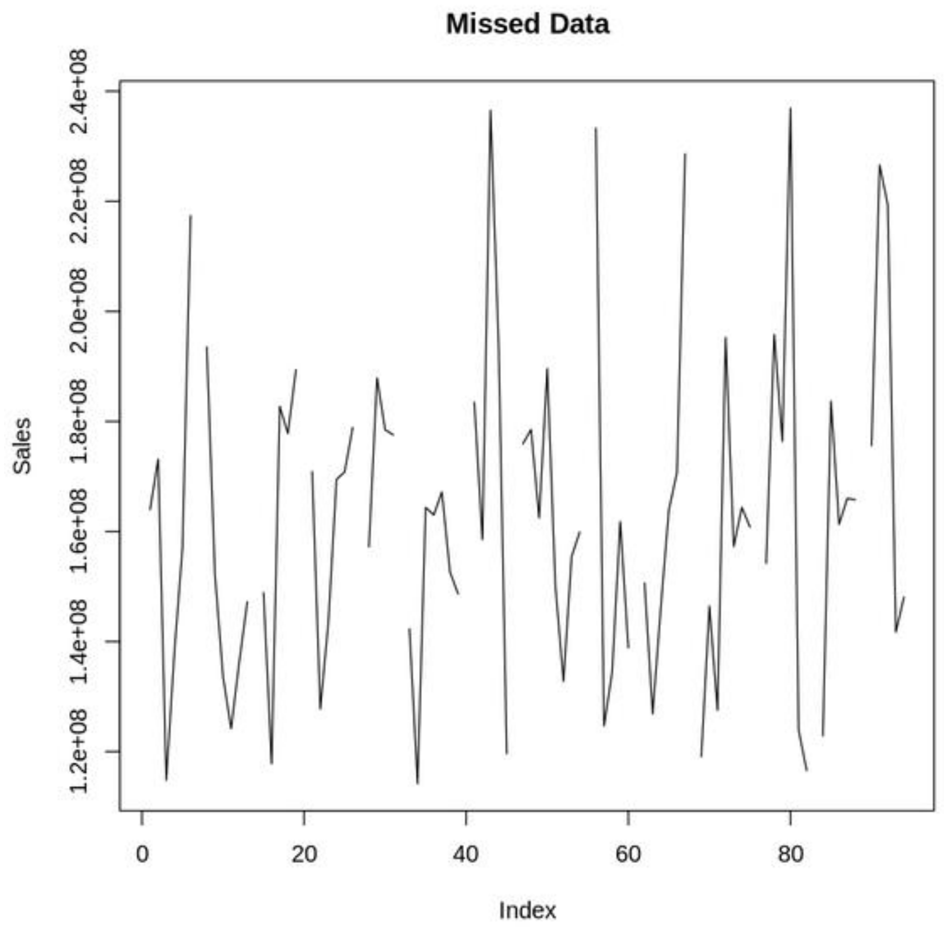
IMPLEMETATION PROCESS:



Pipeline diagram showing the steps in order

II. IMPLEMETATION PROCESS:

Handle missing values

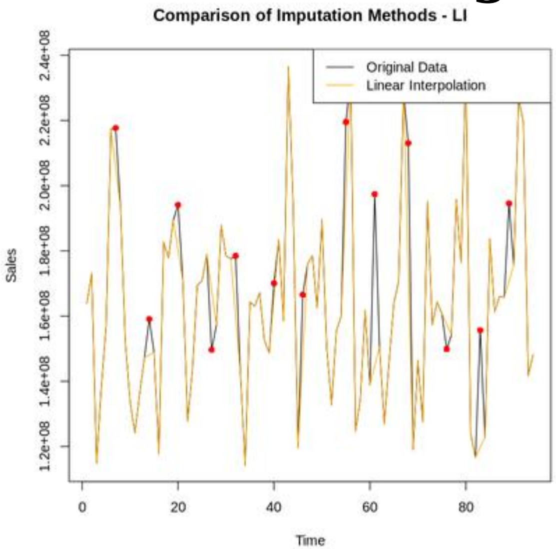


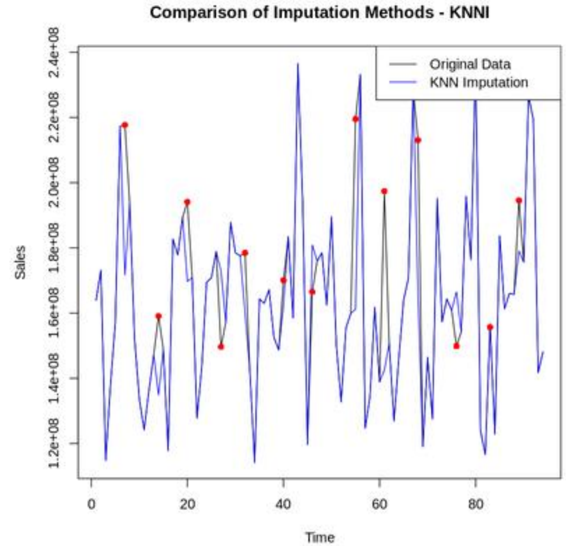
"Missed Data" Chart

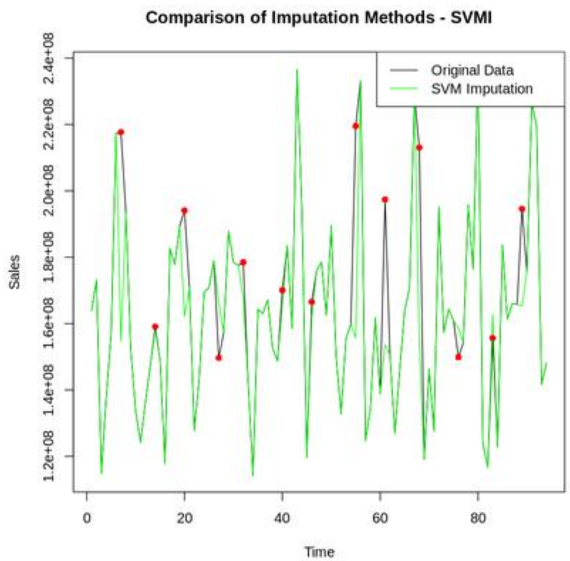
- Gaps show missing (NA) values in Sales.
- Missing points appear uneven and periodic.
- Continuous data needed for ARIMA/ETS/Prophet.
- Visualization aids imputation and forecast stability.
- With missing percentage is 13.83%

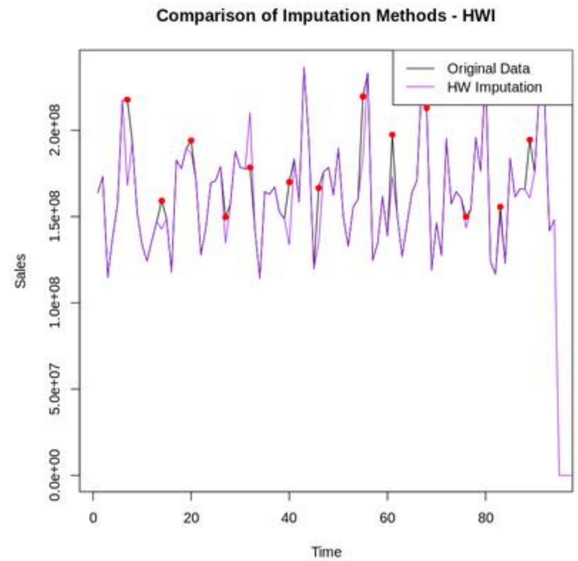
II. IMPLEMETATION PROCESS:

Handle missing values









SVMI and KNNI create excessive volatility, significantly distorting the original trend.

III. RESEARCH RESULT

```
[1] "Comparison of Imputation Methods (RMSE)
                       Method     RMSE
1        Linear Interpolation 25211067
2              KNN Imputation 32357125
3              SVM Imputation 34911481
4 Holt-Winters Based Imputation 26813983
```

Linear Interpolation (LI) and Holt-Winters (HW) are the two most effective imputation methods, achieving significantly lower error metrics compared to KNN Imputation (32.36M) and SVM Imputation (34.91M).

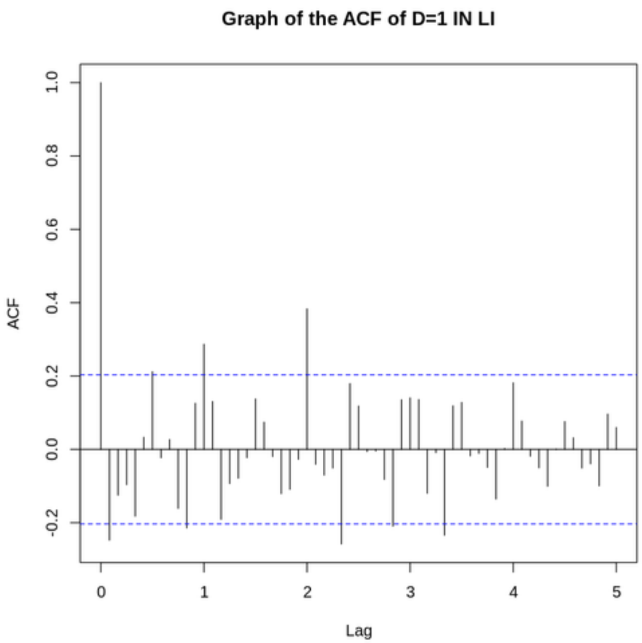
$$Z = \frac{x - \text{mean}}{\text{sd}}$$

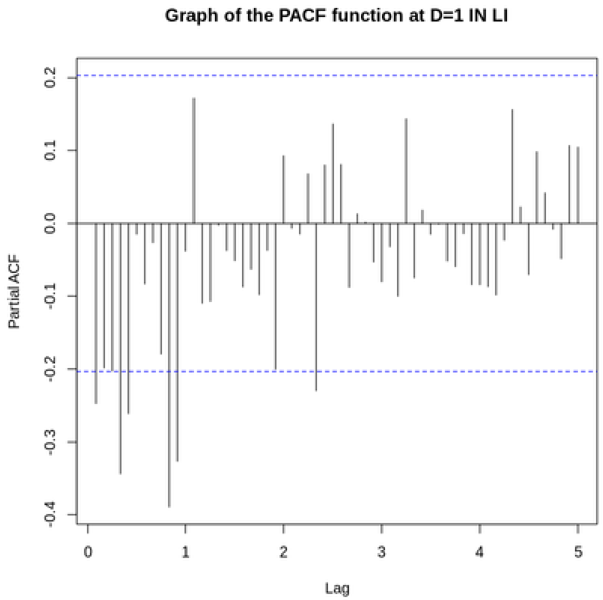
```
[1] "Scaled LI Imputed Data (first 10):"
             [,1]
 [1,]  0.04816805
 [2,]  0.36892683
 [3,] -1.66806523
 [4,] -0.85670943
 [5,] -0.19317029
 [6,]  1.91156707
 [7,]  1.49556575
 [8,]  1.07956443
 [9,] -0.35925208
[10,] -1.00904240
[1] "Scaled HW Imputed Data (first 10):"
             Jan         Feb Mar Apr         May         Jun         Jul
2017                                  0.04857962  0.36347055 -1.63625750
2018 -0.35138731 -0.98929060
             Aug         Sep         Oct         Nov         Dec
2017 -0.83974436 -0.18834377  1.87789027  0.18725752  1.06110800
2018
```

The values have been successfully scaled → suitable for proceeding with building and training the SARIMA forecasting model in the next step

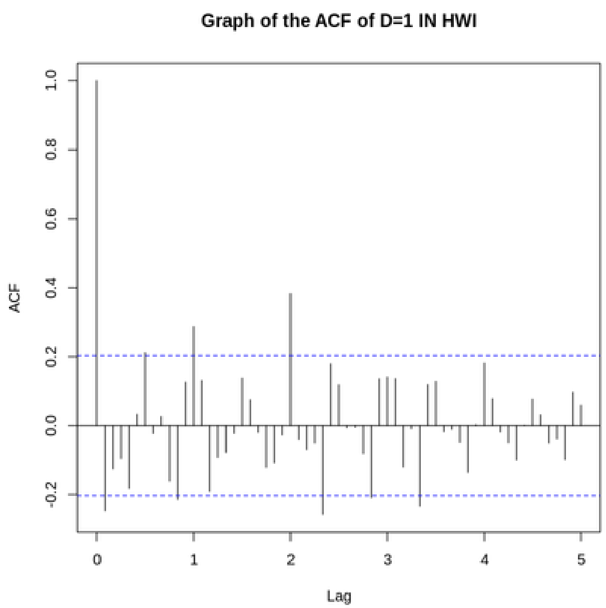
RESEARCH RESULTS

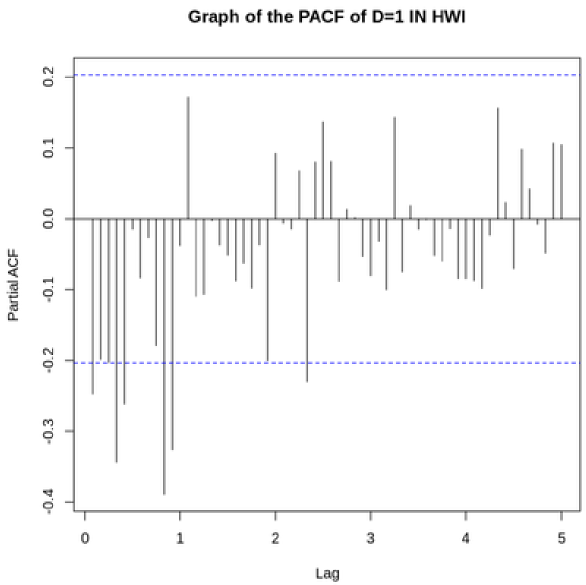
Seasonal Data





LI
Q = 1,2,3
P=1,3

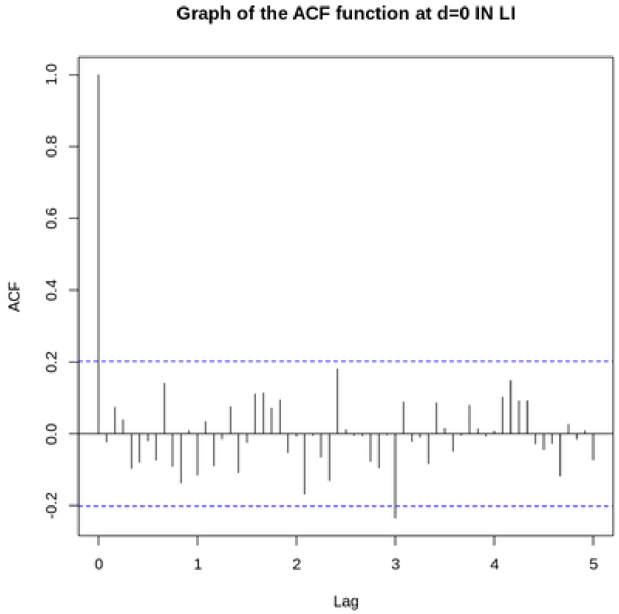


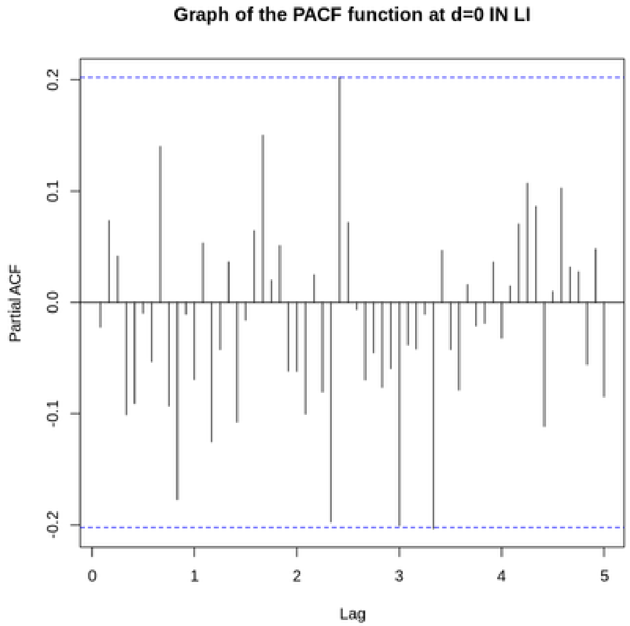


HWI
Q = 1,2,3,4
P = 1,2

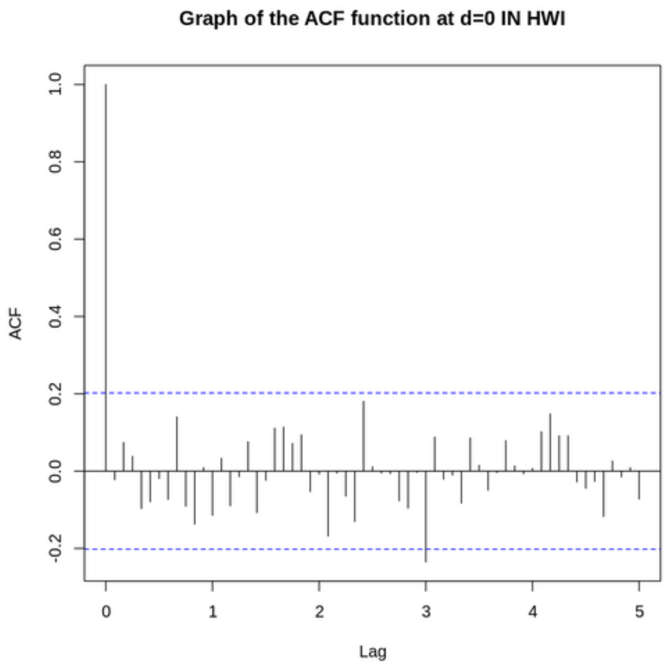
RESEARCH RESULTS

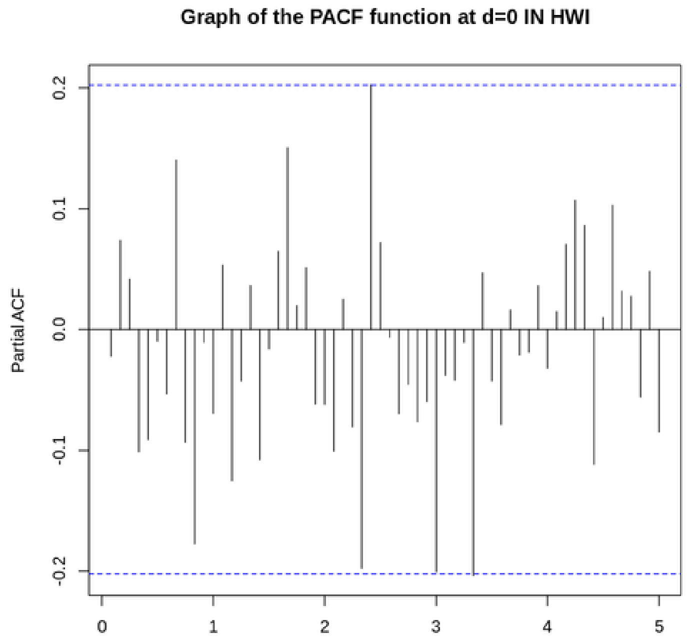
Non-Seasonal Data





LI
q=3
p=4





HWI
q=1,3
p=1,3

RESEARCH RESULTS

LI TIME

A data.frame: 6 × 2

	df	AIC
	<dbl>	<dbl>
MH1	10	201.5680
MH2	11	200.7935
MH3	12	198.9715
MH4	12	201.3156
MH5	13	203.2724
MH6	14	204.6986

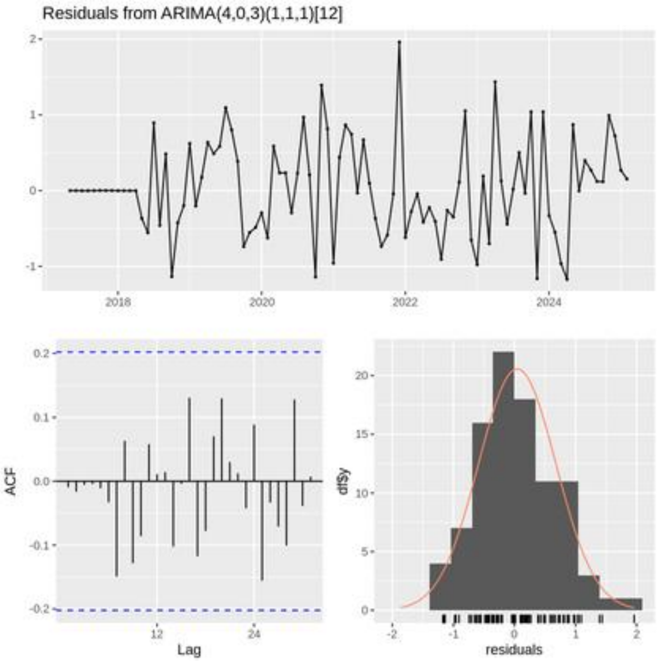
A data.frame: 6 × 2

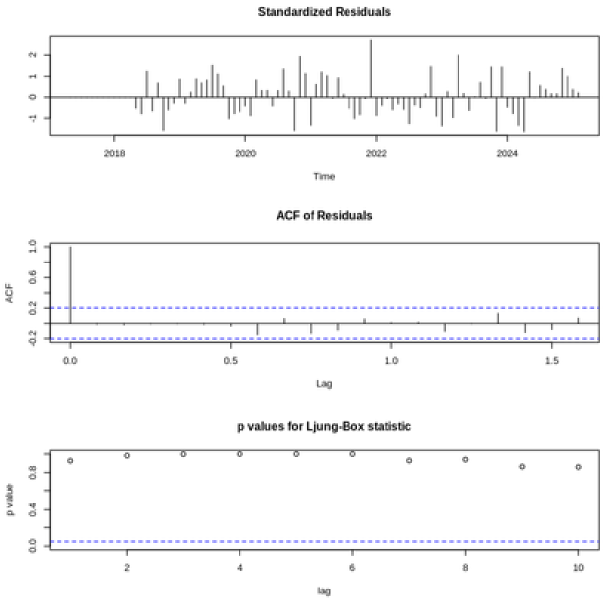
	df	BIC
	<dbl>	<dbl>
MH1	10	225.6352
MH2	11	227.2674
MH3	12	227.8521
MH4	12	230.1962
MH5	13	234.5598
MH6	14	238.3927

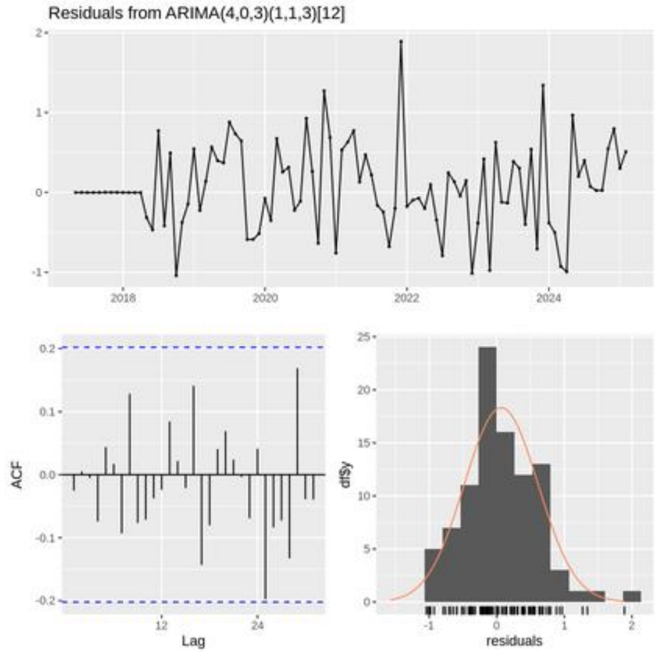
AIC: MH3
BIC: MH1

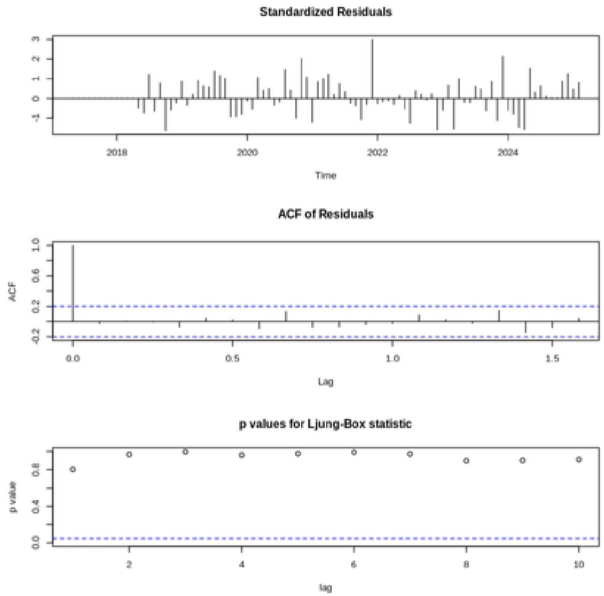
RESEARCH RESULTS

LI TIME





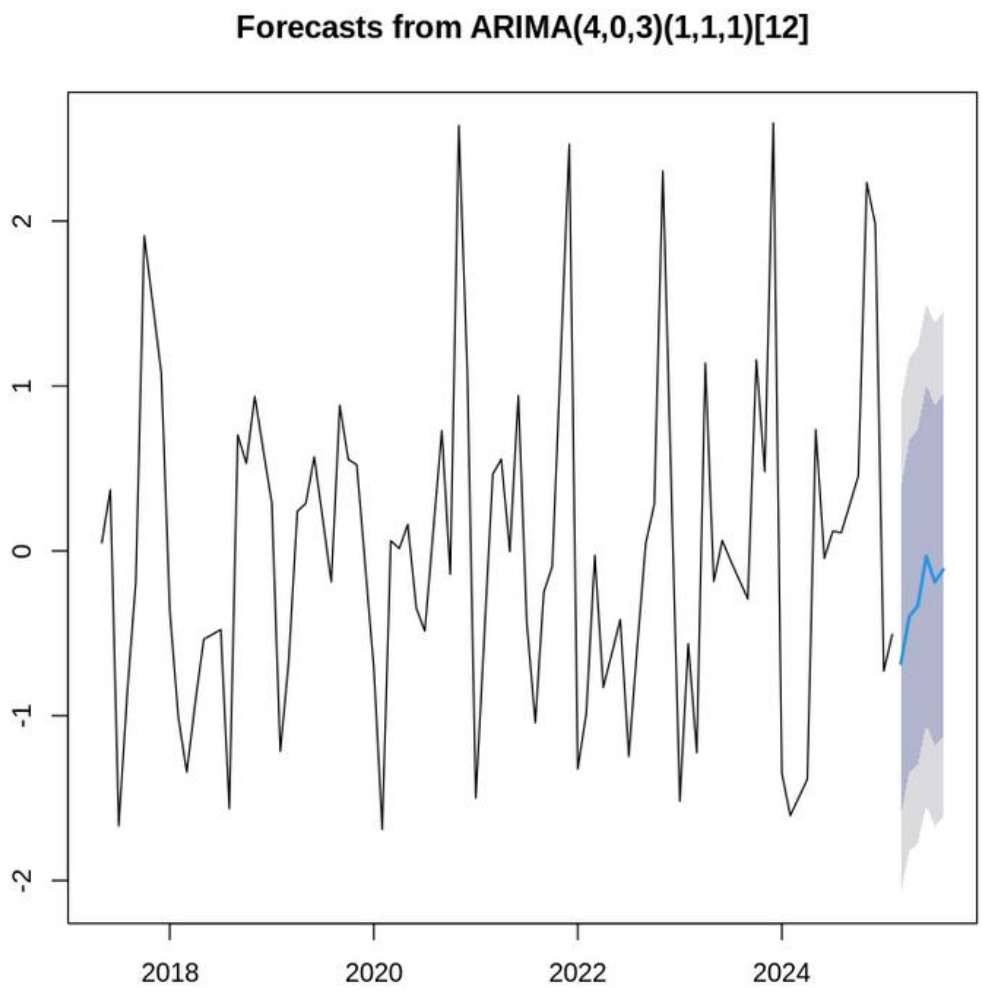




The SARIMA(4,0,3)(1,1,1)[12] model shows a slight superiority

RESEARCH RESULTS

LI TIME



The forecasting model SARIMA(4,0,3)(1,1,1)[12] is capable of capturing the variability and seasonality of the time series.

RESEARCH RESULTS

HWI

A data.frame: 20 × 2

	df	AIC
	<dbl>	<dbl>
HW1	5	170.3287
HW2	6	171.5360
HW3	7	171.0658
HW4	8	172.7788
HW5	6	172.1463
HW6	7	171.9909
HW7	8	173.0185
HW8	9	174.3972
HW9	7	173.8208
HW10	8	174.8631
HW11	9	173.2805
HW12	10	174.6266
HW17	7	173.7520
HW18	8	174.9037
HW19	9	174.5669
HW20	10	175.9248
HW25	9	174.4158
HW26	10	175.1079
HW27	11	172.8686
HW28	12	175.9321

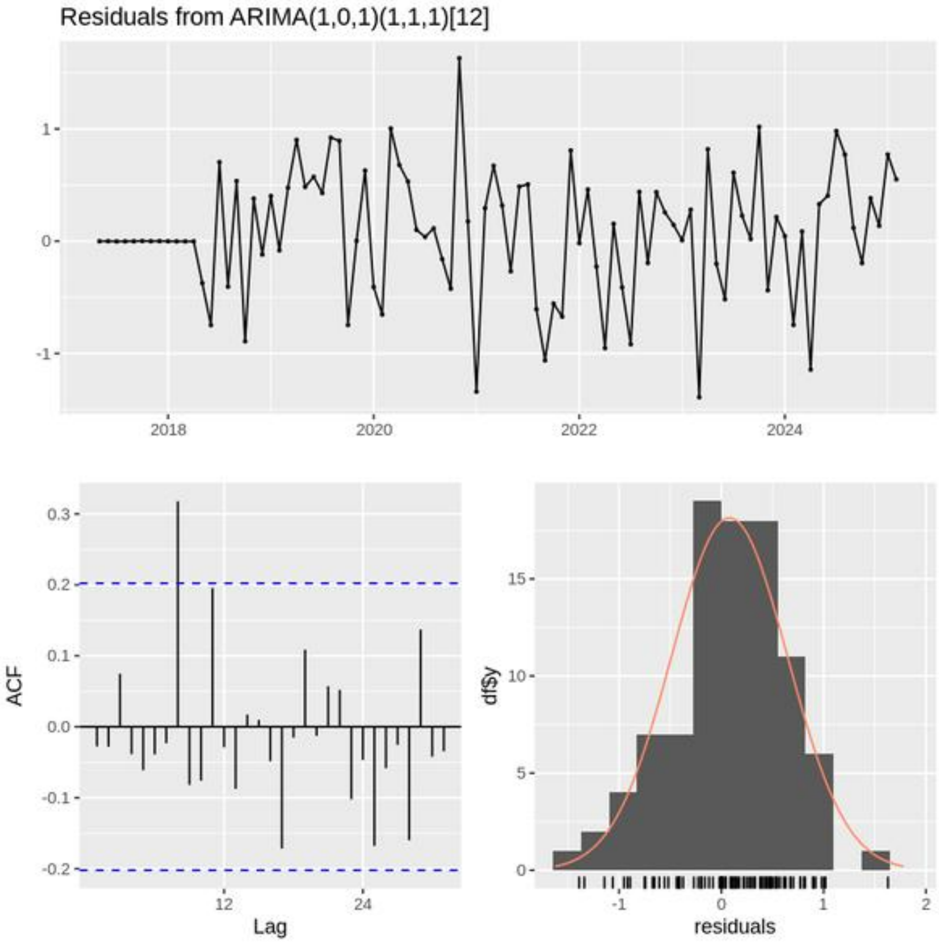
A data.frame: 20 × 2

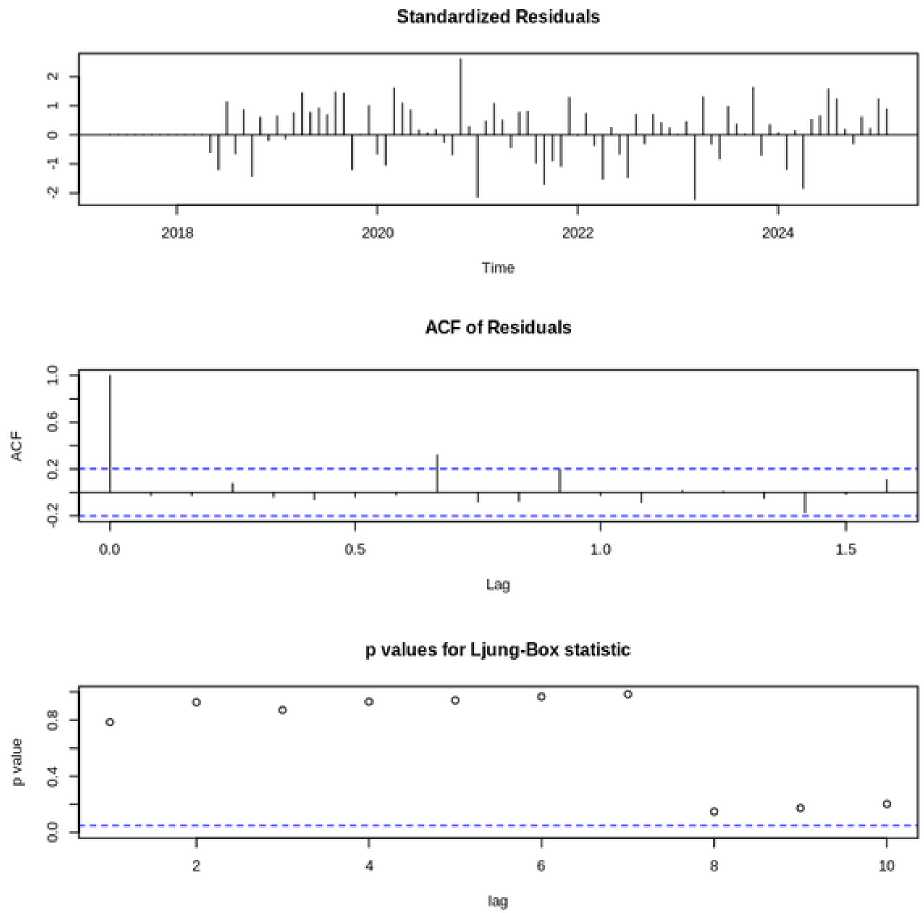
	df	BIC
	<dbl>	<dbl>
HW1	5	182.3623
HW2	6	185.9763
HW3	7	187.9129
HW4	8	192.0326
HW5	6	186.5866
HW6	7	188.8380
HW7	8	192.2723
HW8	9	196.0576
HW9	7	190.6679
HW10	8	194.1169
HW11	9	194.9410
HW12	10	198.6938
HW17	7	190.5990
HW18	8	194.1574
HW19	9	196.2274
HW20	10	199.9920
HW25	9	196.0763
HW26	10	199.1751
HW27	11	199.3425
HW28	12	204.8128

AIC/BIC: HW1

RESEARCH RESULTS

HWI

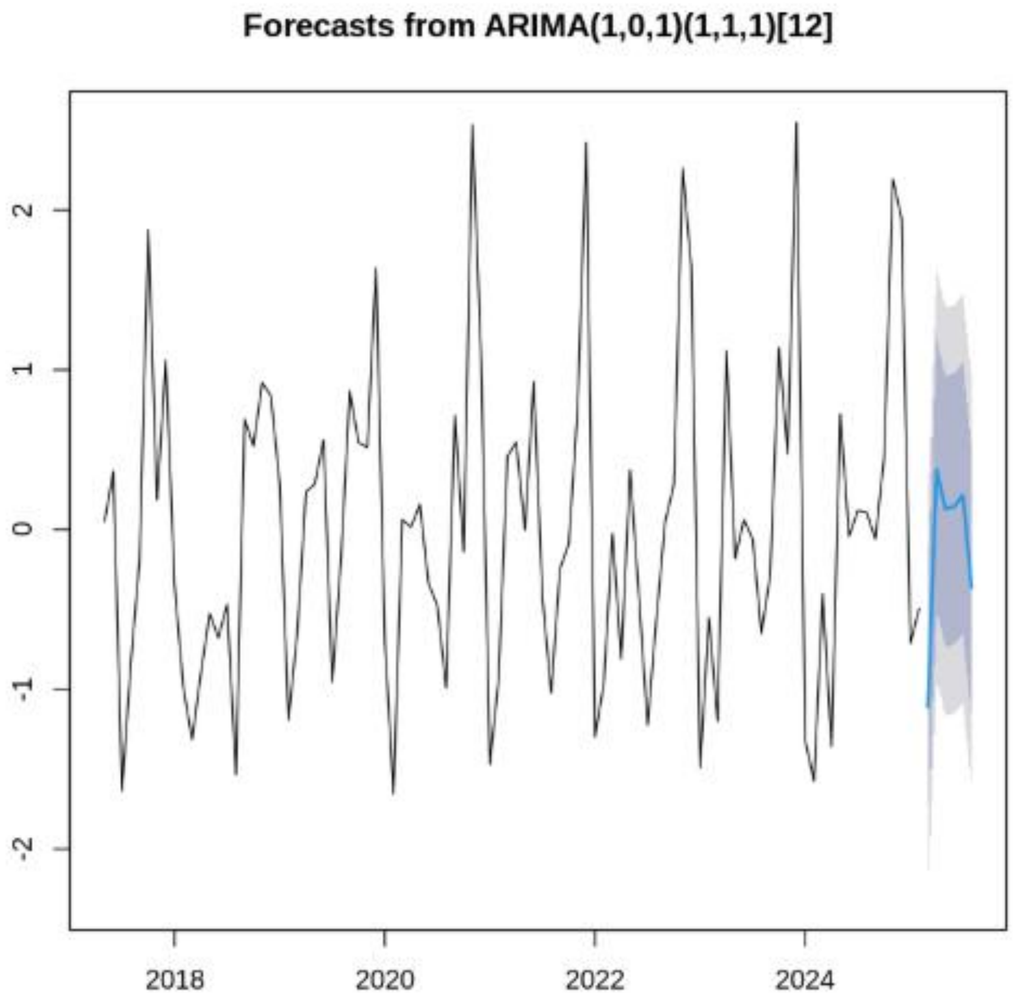




The SARIMA(1,0,1)(1,1,1)[12] model is appropriate and adequate for describing the time series, as the residuals are nearly white noise, indicating that the model has captured most of the data structure.

RESEARCH RESULTS

HWI



The SARIMA(1,0,1)(1,1,1)[12] model provides good short-term forecasts, but the uncertainty increases rapidly.

RESEARCH RESULTS

```
[1] "Comparison of Actual and Forecast Values (Starting March 2025):"
    Datetime Actual_Values Forecast_LI Forecast_HWI
1 2025-03-01     145399492   143001507    130103418
2 2025-04-01     157758923   151196176    173517020
3 2025-05-01     176006666   153055016    166310141
4 2025-06-01     182082611   161647008    166728412
5 2025-07-01     205509243   157160697    168849748
6 2025-08-01     236514347   159345408    151897607
```

- Both models failed to track the strong, accelerating growth in Actual_Values (especially May-Aug).
- Forecast_LI was too conservative (underestimated the trend).
- Forecast_HWI was volatile, with large errors (overestimated in April, failed at the August peak).

RESEARCH RESULTS

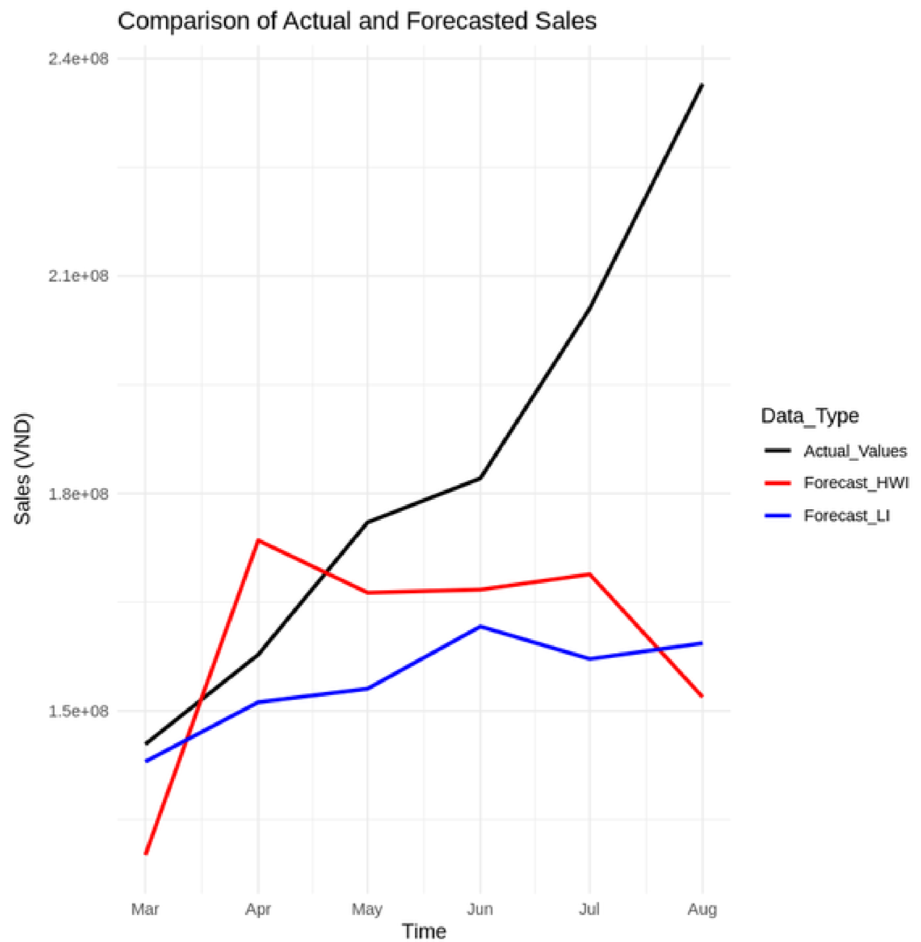
Table: Comparative Analysis of Revenue Forecasting Models

Models	MAE	RMSE	MAPE (%)
Linear Interpolation Forecast	39340054	29644245	14.37108
Holt-Winters Based Imputation Forecast	39403834	29563522	14.67757

- Performance Comparison: The Linear Interpolation (LI) method demonstrated superior forecasting performance compared to Holt-Winters (HWI).
- Error Metrics: LI achieved a lower RMSE and MAPE (14.37%) compared to HWI.

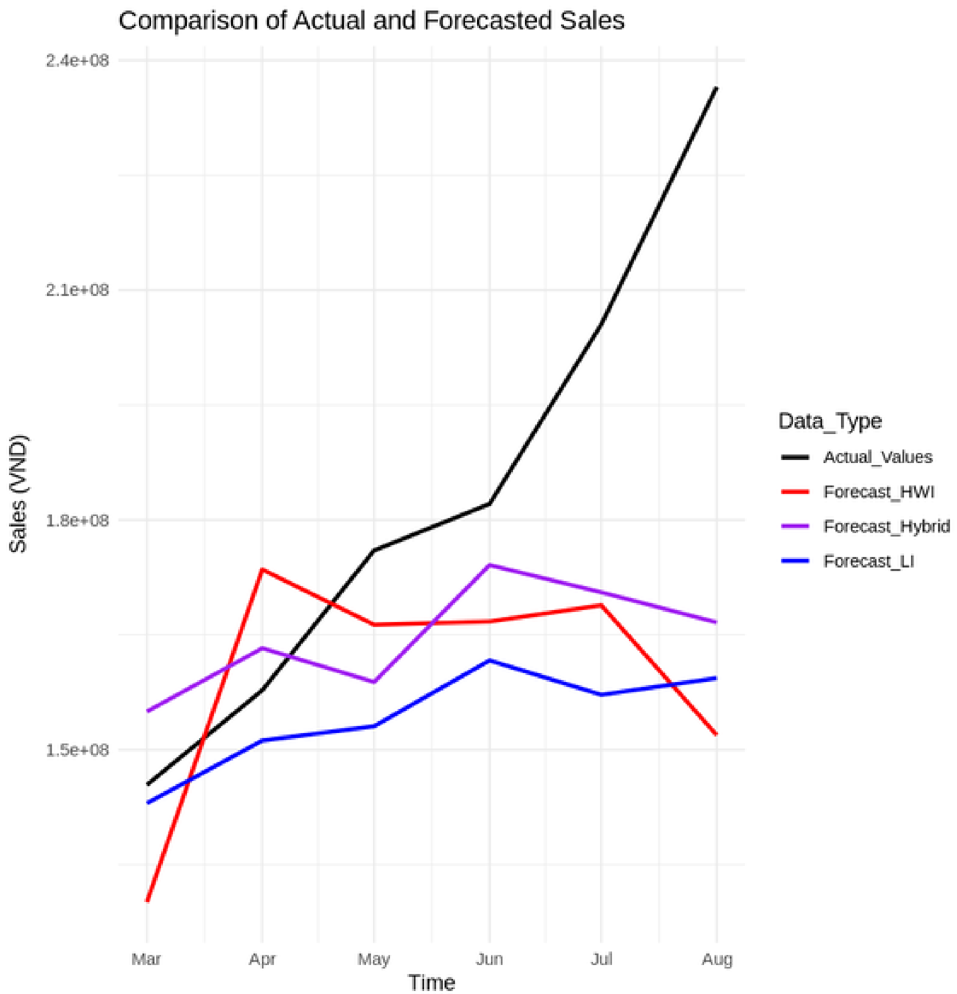
→ Based on the error metrics, LI is the most suitable method for imputing missing values, ensuring the continuity of the revenue time series data.

RESEARCH RESULTS



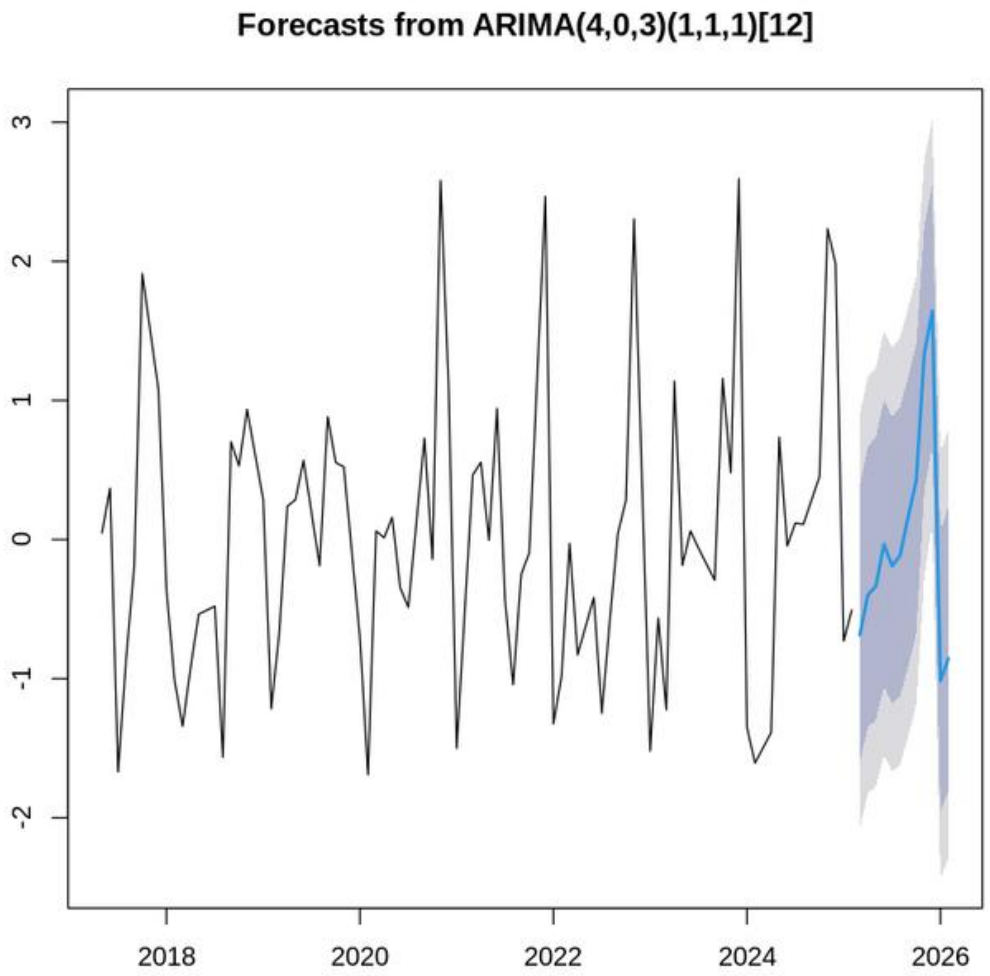
- Both the LI and HWI forecast models significantly underestimate this growth rate, but LI exhibits a smoother and more stable trend.
- Based on the stability of the trend and associated error metrics, LI is the more optimal model compared to HWI for initial estimation.

RESEARCH RESULTS



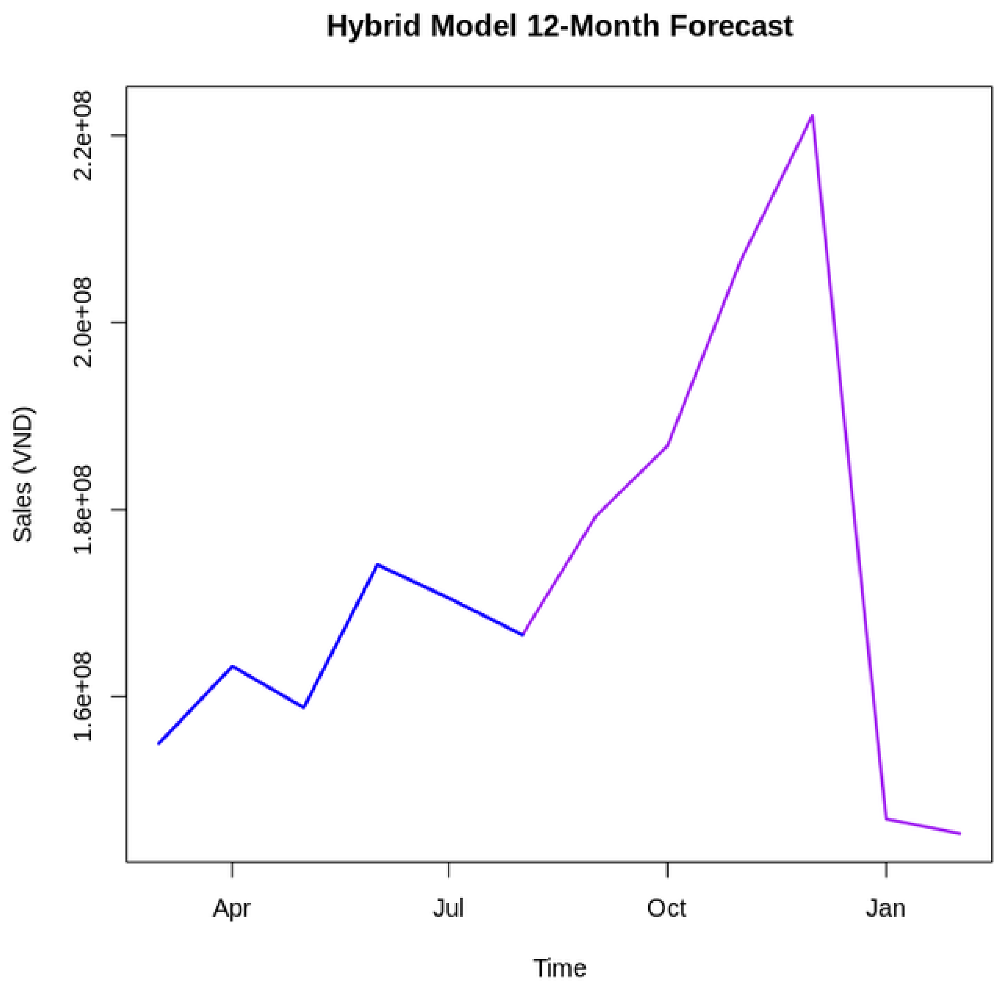
- Actual sales show a strong upward trend
- The Forecast_Hybrid model is the most accurate, closely tracking the general trend, significantly outperforming Forecast_HWI and Forecast_LI.

RESEARCH RESULTS



- The SARIMA forecast chart clearly demonstrates the seasonality of the time series.
- The forecast reaches its growth peak at the end of 2025.

RESEARCH RESULTS



The Hybrid Model's 12-month forecast shows revenue peaking highest in December, followed by a sharp drop in January, confirming the pronounced seasonality of the market.

CONCLUSION & STRATEGIC RECOMMENDATIONS:

Conclusion:

- Linear Interpolation (LI) is the most suitable method to fill missing data, ensuring the continuity and actual trend of the revenue time series.
- The Hybrid SARIMA–XGBoost model demonstrated high accuracy (low MAE, RMSE), precisely capturing the growth trend and seasonal characteristics of the ceramics revenue.
- Diagnostic tests confirmed that the residuals are random and show no autocorrelation, ensuring the reliability of the forecasts.
- The model serves as a direct support tool for ceramics shops in business decision-making (optimizing production, distribution, and inventory) based on seasonal cycles.

CONCLUSION & STRATEGIC RECOMMENDATIONS:

Strategy:

1. **Production – Supply:** Sufficient stock and flexibility against market demand → Increase capacity by 20-30% and early inventory stocking of raw materials (before Q3)
2. **Human Resources – Operations:** Maintain high productivity while minimizing labor costs during off-peak seasons → Seasonal labor recruitment and training, coupled with the application of automation, to limit human dependence.
3. **Marketing – Sales:** Expand market share and capitalize on the year-end growth peak → Increase Marketing budget by 40% for Lunar New Year and year-end campaigns (Q3, Q4).
4. **Financial – Risk:** Ensure stable cash flow and mitigate risks following the peak season → Increase cash reserves by 10-15% (pre-Dec) and establish short-term credit lines

THANKS FOR LISTENING!