**TRUCK SALES TIME SERIES ANALYSIS AND PREDICTION**

*Author: Nguyen Ngoc Tue Minh*

*2024*

**Table of Contents**

[Table of Figures 3](#_Toc178084577)

[Introduction and project purposes 4](#_Toc178084578)

[Dataset overview 4](#_Toc178084579)

[Time series analysis 4](#_Toc178084580)

[Identifying patterns and decomposition 5](#_Toc178084581)

[Stationarity check and preliminary visualizations 5](#_Toc178084582)

[Data modelling 7](#_Toc178084583)

[Compare models' performance 10](#_Toc178084584)

[Evaluate model performance with Time Series Cross Validation 11](#_Toc178084585)

[Conclusion 11](#_Toc178084586)

[References 11](#_Toc178084587)

# Table of Figures

[Figure 1 4](#_Toc163127141)

[Figure 2 4](#_Toc163127142)

[Figure 3 5](#_Toc163127143)

[Figure 4: ADF test on original data 5](#_Toc163127144)

[Figure 5 6](#_Toc163127145)

[Figure 6: ADF test on 1st non-seasonal difference 6](#_Toc163127146)

[Figure 7: ADF test on 2nd non-seasonal difference 6](#_Toc163127147)

[Figure 8 7](#_Toc163127148)

[Figure 9: ADF test on 1st seasonal difference 7](#_Toc163127149)

[Figure 10: SARIMA(0, 1, 1)(0, 1, 1, 12) result 8](#_Toc163127150)

[Figure 11: SARIMA(0, 1, 1)(0, 1, 1, 12) result 8](#_Toc163127151)

[Figure 12: SARIMA(0, 1, 1)(0, 1, 1, 12) result 9](#_Toc163127152)

[Figure 13 10](#_Toc163127153)

[Figure 14 10](#_Toc163127154)

[Figure 15 11](file:///D:\Master%20in%20Financial%20Analytics\8.%20Applied%20financial%20analytics\CA2\Time%20series%20analysis%20-%20Truck%20sales%20-%20Report%20-v2.docx#_Toc163127155)

[Figure 16 11](file:///D:\Master%20in%20Financial%20Analytics\8.%20Applied%20financial%20analytics\CA2\Time%20series%20analysis%20-%20Truck%20sales%20-%20Report%20-v2.docx#_Toc163127156)

# Introduction and project purposes

This project use time series analysis techniques to analyse monthly sales of a truck company over a period of 12 years and build time series models to predict sales for the last 12 months.

# Dataset overview

[This dataset](https://www.kaggle.com/datasets/ddosad/dummy-truck-sales-for-time-series) from Kaggle has 144 rows containing information on monthly sales from 2003 to 2014. Sales range went from 152 to 958 units. 50% of monthly sales stayed between 273 and 560. There is no missing values and outliers in this dataset.

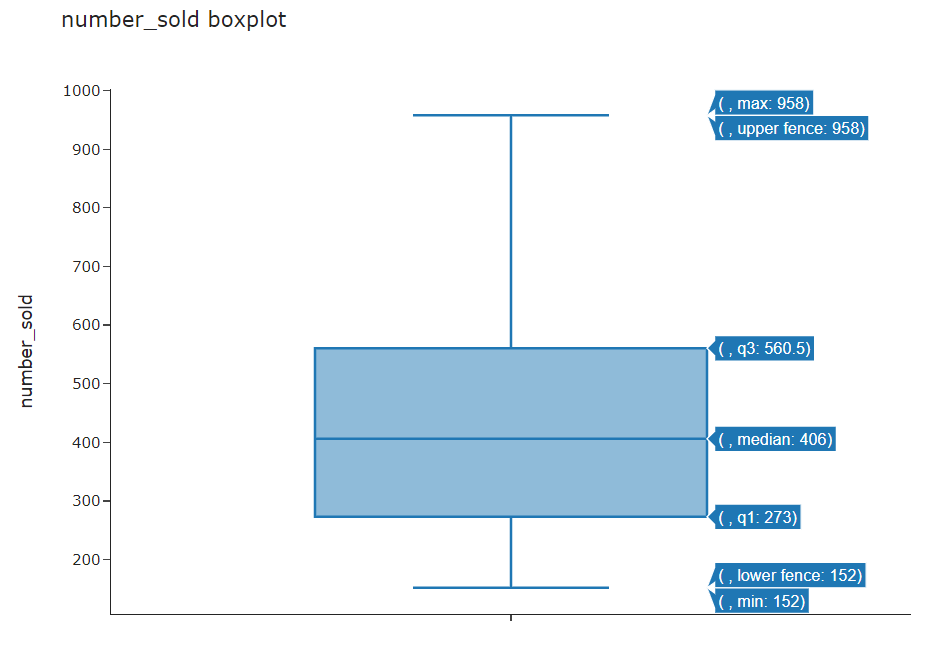


Figure 1

# Time series analysis

The line chart illustrates a yearly pattern. This pattern shows the sales gradually climbing and peaking toward the end of the year but ending with a sudden reduction. This peak occurs in the months of July and August.

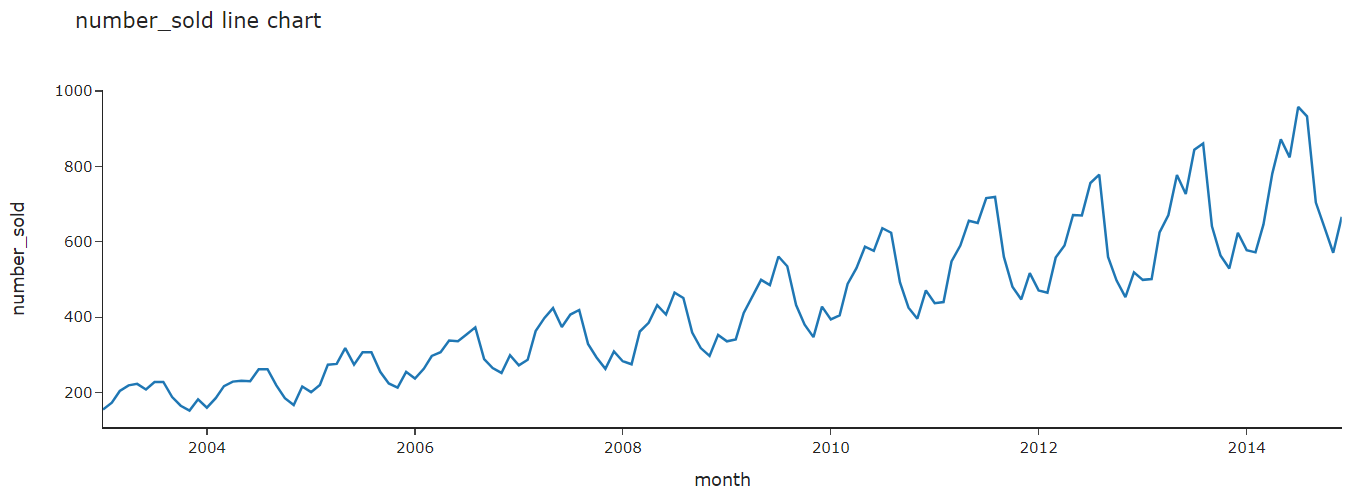


Figure 2

## Identifying patterns and decomposition

The time series is multiplicative because the magnitude of the seasonality component is dependent of the trend (Sigmundo Preissler Jr, 2018). The line plot above shows there is a 12-month-seasonality. Decomposition as below:

* The original data reflects the combined effects of trend, seasonality, and noise.
* The trend indicates a consistent long-term increase.
* Seasonality shows regular peaks and drops every year.
* Residuals mostly shows small random changes, but sometimes there are bigger jumps or drops that aren't explained by the model

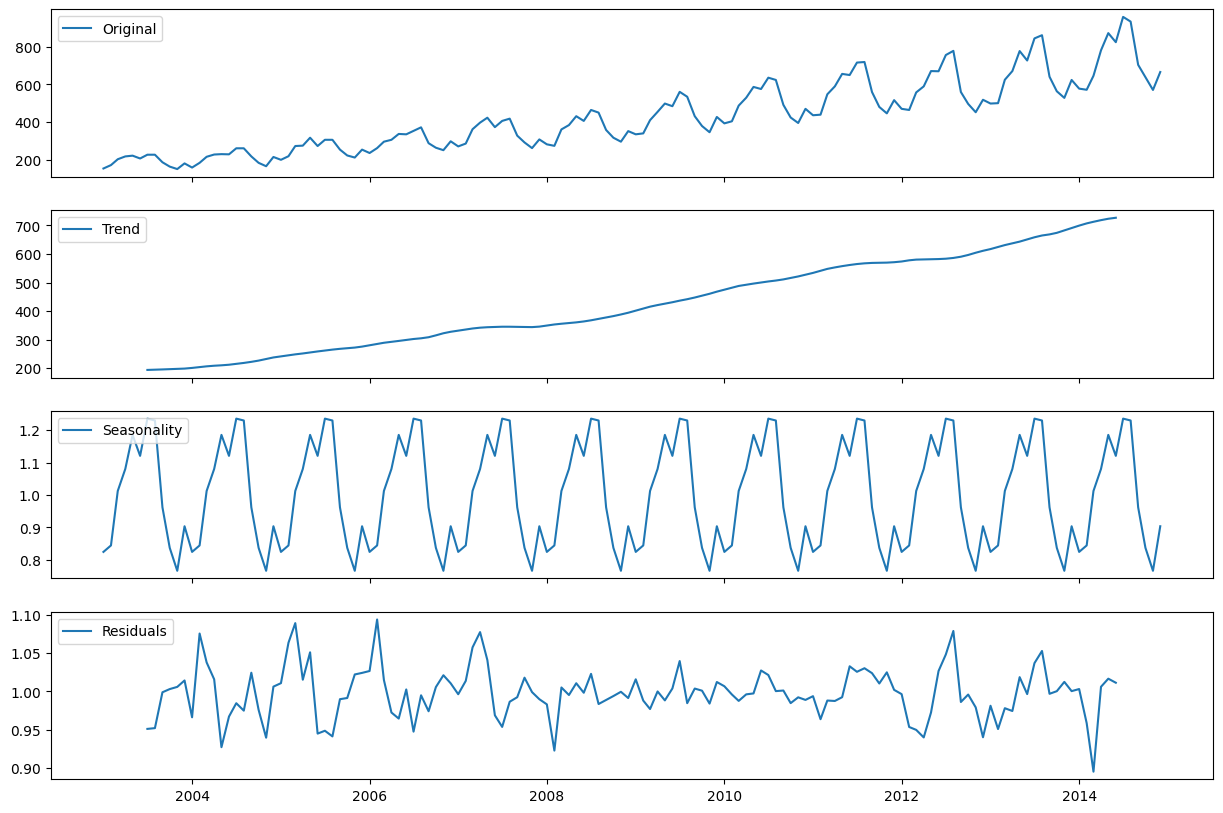


Figure 3

## Stationarity check and preliminary visualizations

The ADF test on the original series shows a high p-value (0.995), indicating non-stationarity. Therefore, we need to differencing to make it stationary.

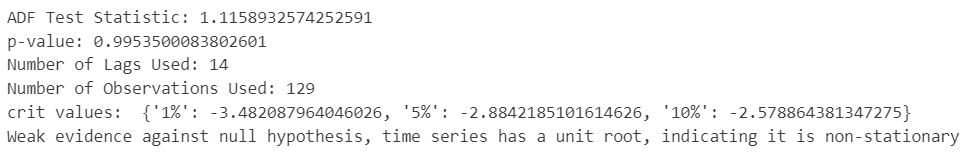


Figure 4: ADF test on original data

**Order of non-seasonal differencing**

ACF and PACF plot show a trend and uncaptured seasonality after the first difference. Second differencing shows signs of over-differenced by many negative lags.

ADF test for first differencing doesn't fully stabilize the series (p-value > 0.05) but second differencing achieved stationary (p-value < 0.05), but possible over-differencing because of many negative spikes.

Given these observations and the clear seasonality in the data, we tried seasonal differencing to avoid too much non-seasonal differencing.

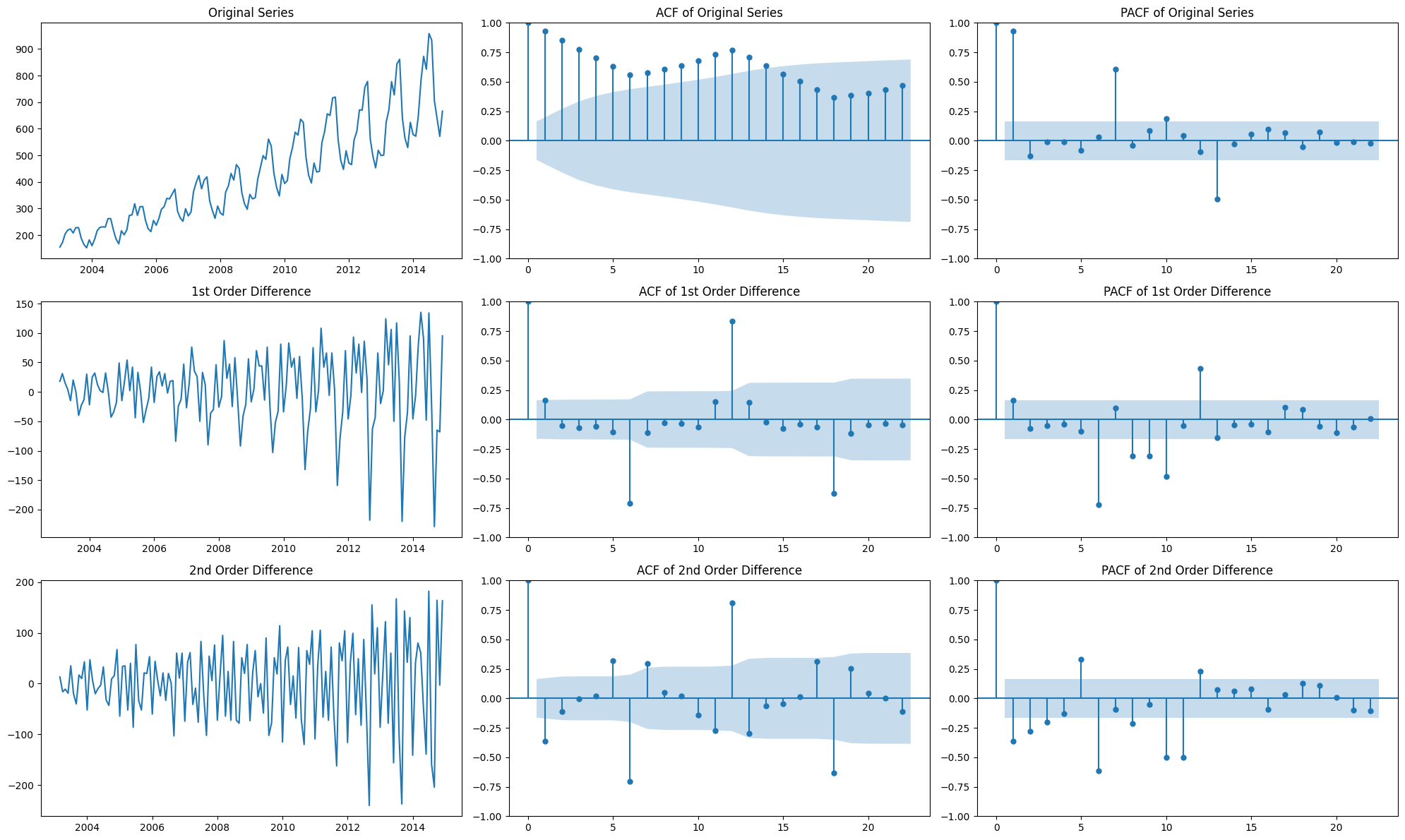


Figure 5

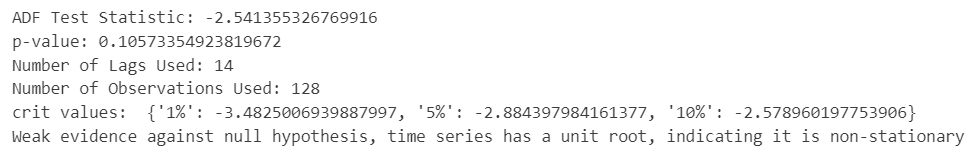


Figure 6: ADF test on 1st non-seasonal difference

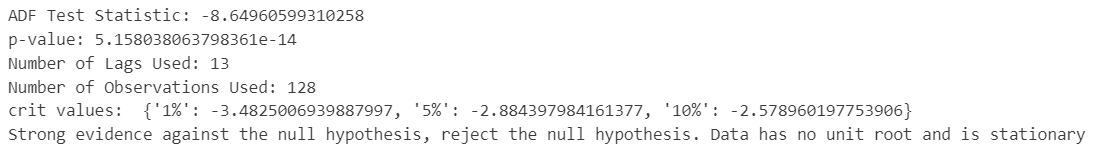


Figure 7: ADF test on 2nd non-seasonal difference

**Order of seasonal differencing**

The 1st order seasonal difference removes some of the trend and seasonality but still shows short-term correlation. Result from ADF test shows stationarity.

The 2nd order seasonal difference may be too much, as indicated by the ACF and PACF patterns, which no longer show a clear seasonal pattern and suggest possible over-differencing.

Based on these observations, we modelled the dataset with a 1st order seasonal difference.

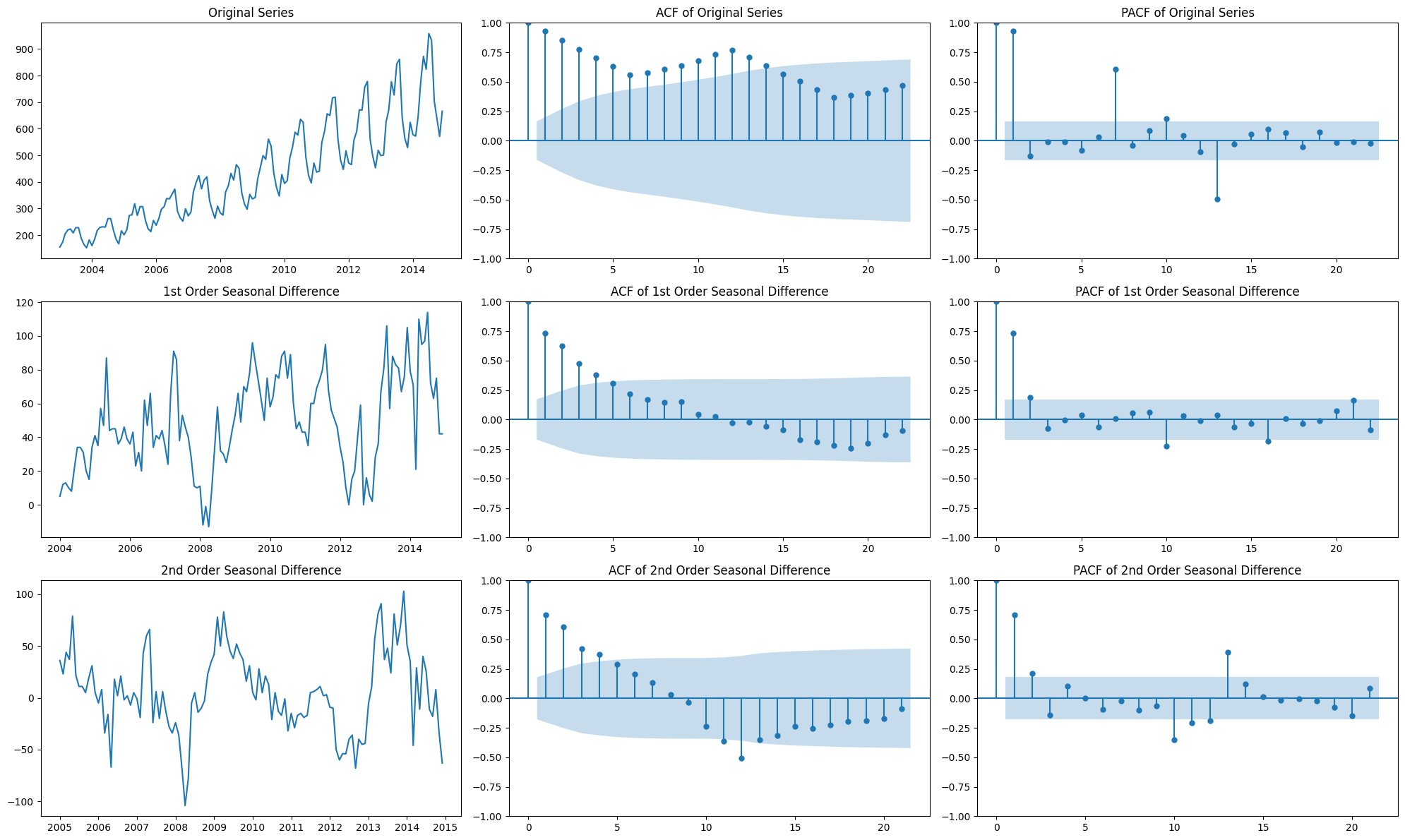


Figure 8

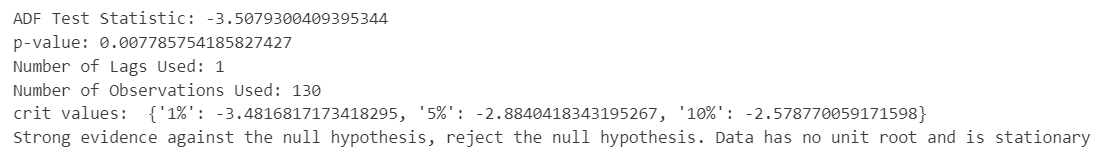


Figure 9: ADF test on 1st seasonal difference

# Data modelling

We created five models (manual\_ARIMA, manual\_SARIMA, Facebook Prophet, auto\_ARIMA, auto\_SARIMA) to forecast sales for the last 12 months, following these steps for each model:

* Analyzed the train dataset with ACF and PACF plots.
* Conducted ADF and KPSS tests for stationarity.
* Applied logarithmic transformation if initial differences didn't stabilize the time series.
* Identified the appropriate differencing and model terms (AR, MA) using ACF and PACF.
* Fitted the model to the training set and forecasted the test dataset of 12 months.
* Calculated accuracy metrics (MAE, MAPE, MSE, R2) and plotted residual diagnostics.

We compared their performance and selected SARIMA(0,1,1)(0,1,1,12) as the top performer. Then we evaluated this model via time series cross-validation to assess its reliability over time.

Due to the assignment’s word-limitation, we only presented the result of the best model.

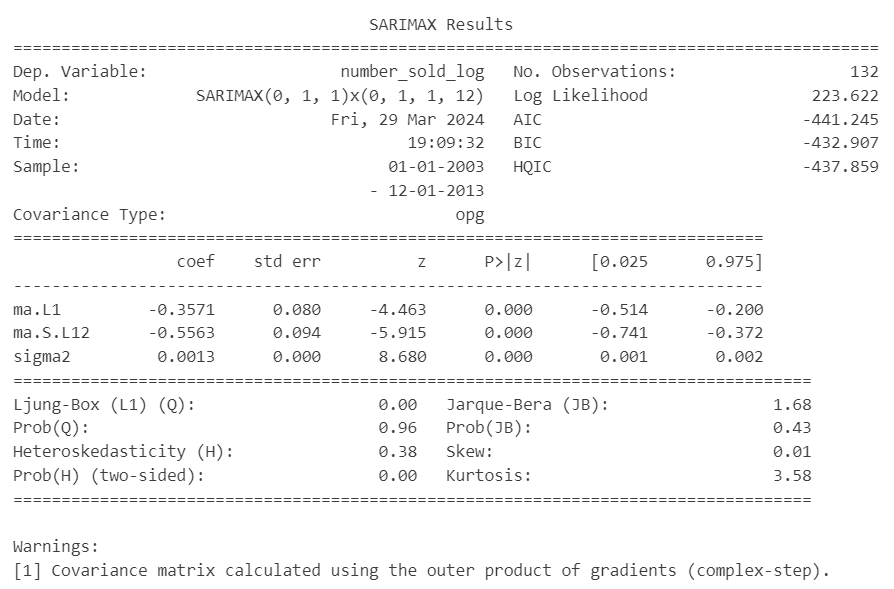


Figure 10: SARIMA(0, 1, 1)(0, 1, 1, 12) result

Evaluation measures:

* MAE: 20.14 units off on average from actual sales.
* MAPE: approximately 2.90% average deviation from actual values, indicating high accuracy.
* MSE: roughly 794.19, highlighting typical squared errors in predictions.
* R2: 95.63% variance in truck sales explained by this model, demonstrating excellent explanatory power.

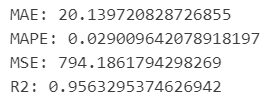


Figure 11: SARIMA(0, 1, 1)(0, 1, 1, 12) result

The diagnostic plots show the model effectively captures the time series trends:

* Standardized\_Residual: Lack of patterns in residuals implies a good model fit.
* Histogram\_plus\_Estimated Density: The close match with the normal distribution suggests accurate residual modelling.
* Normal\_Q-Q\_Plot: Some deviations at the ends, but the dots follow the line quite closely, supports residual normality.
* Correlogram: Autocorrelation is within acceptable limits, indicating no significant residual dependencies.

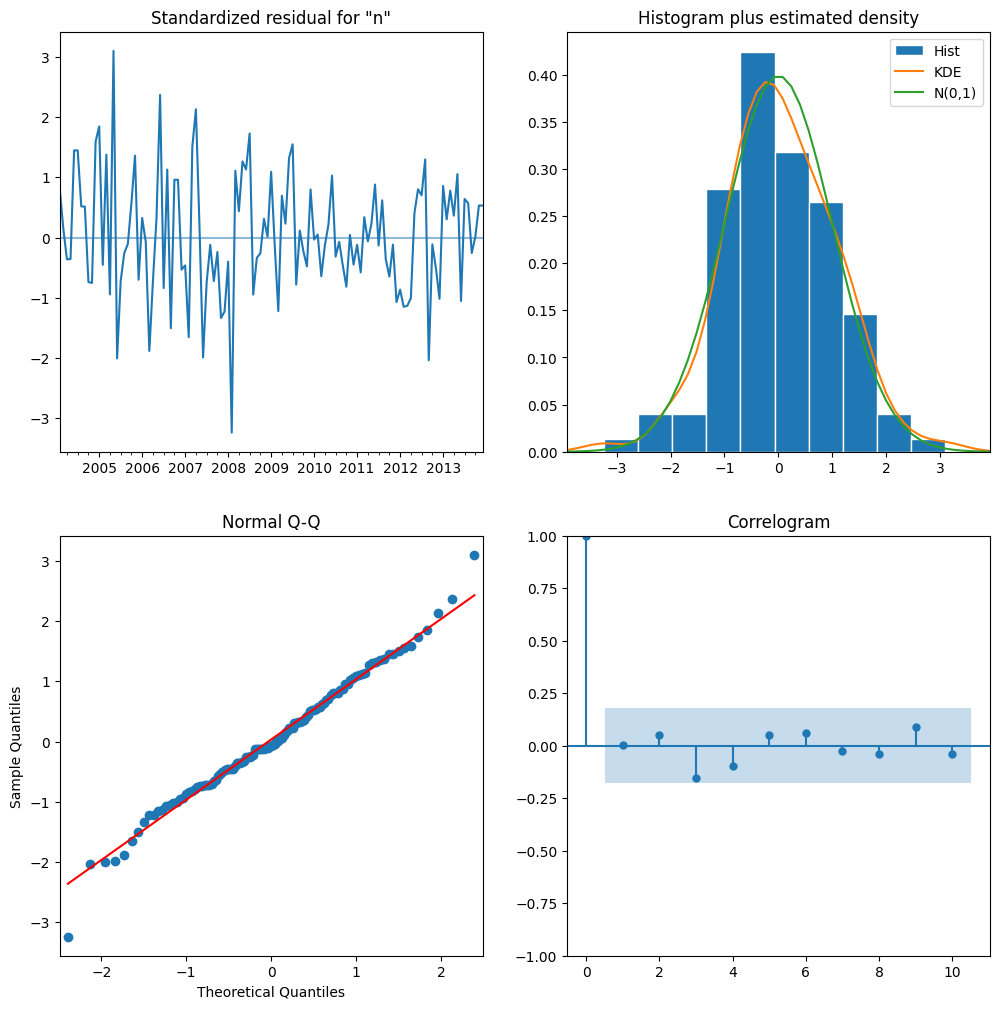


Figure 12: SARIMA(0, 1, 1)(0, 1, 1, 12) result

# Compare models' performance

Models' performance comparison:

* Manual\_ARIMA(0,1,0): has a negative R2 of -0.60, indicating worse performance than a naive average model, and an AIC of -161.65, suggesting poor fit.
* Manual\_SARIMA(0,1,0)(0,1,0,12): shows improvement with an R2 of 0.77, explaining about 77% of variability, and a better AIC of -398.61.
* Manual\_Prophet: R2 of 0.64, indicating reasonable fit. Performing better than manual ARIMA but not as well as manual SARIMA, possibly due to its preference for additive models.
* Auto\_ARIMA(4,0,4): performs poorly with the lowest R2 of -0.76 and an AIC of -182.95, indicating a less accurate model than simple averages.
* Auto\_SARIMA(0,1,1)(0,1,1,12): outperforms all with the highest R2 of 0.96, showing excellent model accuracy, and the lowest AIC of -441.24, indicating **the best fit**.

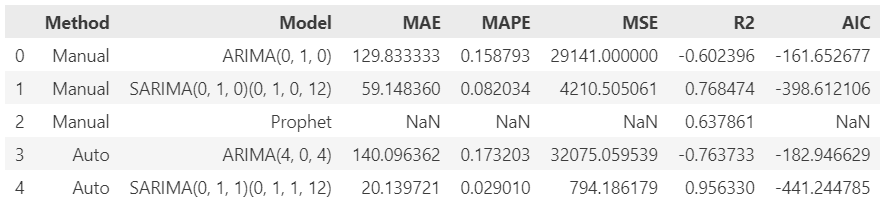


Figure 13

The graph shows SARIMA(0,1,1)(0,1,1,12) matches the closest, indicating the highest accuracy.

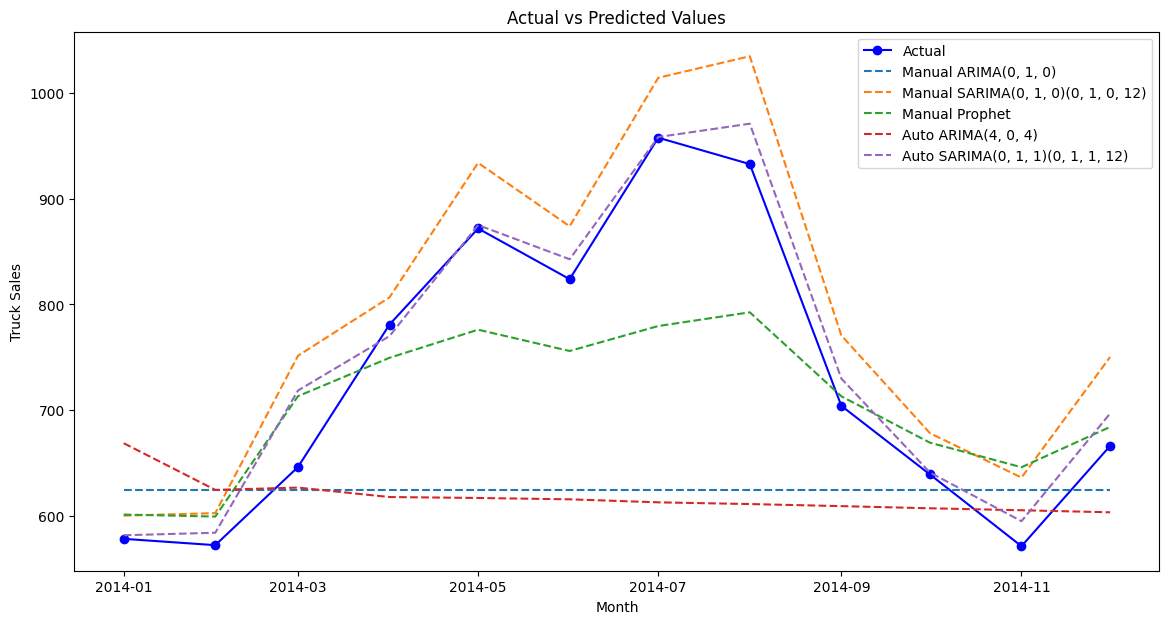


Figure 14

# Evaluate model performance with Time Series Cross Validation

Time series cross-validation helps to verify model performance on unseen data, reducing overfitting risk, and confirm the model's reliability over different periods (Shrivastava, 2020).

The results suggest that SARIMA(0,1,1)(0,1,1,12) generally performs well on the given data.

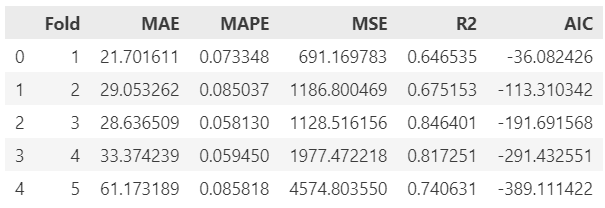
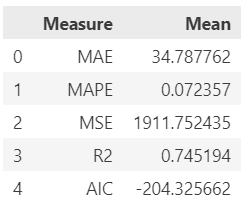
* R2 averages 0.745, lower than the single split's 0.956, but still shows a strong fit.
* AIC averages -204.33, less negative than the single split's -441.24, indicating a less optimal fit compared to the single split model.
* MAE, MAPE and MSE average 34.79, 7.24% and 1911.75 respectively, all are higher than the single split of 20.14, 2.90% and 794.18, indicating larger errors across data subsets.

Figure 15

Figure 16

# Conclusion

We applied time series decomposition, stationarity tests, and forecasting models to predict sales data. Cross-validation shows the SARIMA(0,1,1)(0,1,1,12) model fits well but with higher errors (MAE, MAPE, MSE) than in single-split testing. This suggests the model's general performance can vary with different data segments, highlighting cross-validation’s importance in assessing the model’s predictive strength across time.

# References

Hyndman, R. J. & Athanasopoulos, G., n.d. *Forecasting: Principles and Practice.* [Online]   
Available at: https://otexts.com/fpp2/stationarity.html  
[Accessed 29 March 2024].

Shrivastava, S., 2020. *Cross Validation in Time Series.* [Online]   
Available at: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4  
[Accessed 29 March 2024].

Sigmundo Preissler Jr, P., 2018. *Seasonality in Python: additive or multiplicative model?.* [Online]   
Available at: https://sigmundojr.medium.com/seasonality-in-python-additive-or-multiplicative-model-d4b9cf1f48a7  
[Accessed 29 March 2024].