Retail Sales Time Series Analysis Report

Period: January 2000 to December 2024

1. Introduction

This group project aims to apply key time series analysis techniques covered in the course to a real-world dataset containing monthly Retail Sales Index data from January 2000 to December 2024. The main objective is to explore, analyze, model, and forecast the retail sales index to support future planning decisions.

The analysis involves several critical steps, including visualizing the data through a time plot, identifying trends and seasonal patterns, performing both additive and multiplicative decompositions, and extracting the seasonally adjusted series from the preferred model. The stationarity of the series will be assessed using autocorrelation (ACF) and partial autocorrelation (PACF) plots, followed by differencing when it shows an non-stationary pattern.

A range of models—AR, MA, ARMA, and ARIMA—will be fitted and compared based on statistical criteria such as AIC and residual diagnostics. Finally, the best-fitting model will be used to forecast the next 12 months of retail sales, along with confidence intervals. The entire analysis will be conducted using the R programming language, and all modeling decisions will be clearly justified to demonstrate a solid understanding of time series concepts.

2. Time Plot and Description

The data is first converted into a time series object and plotted below:

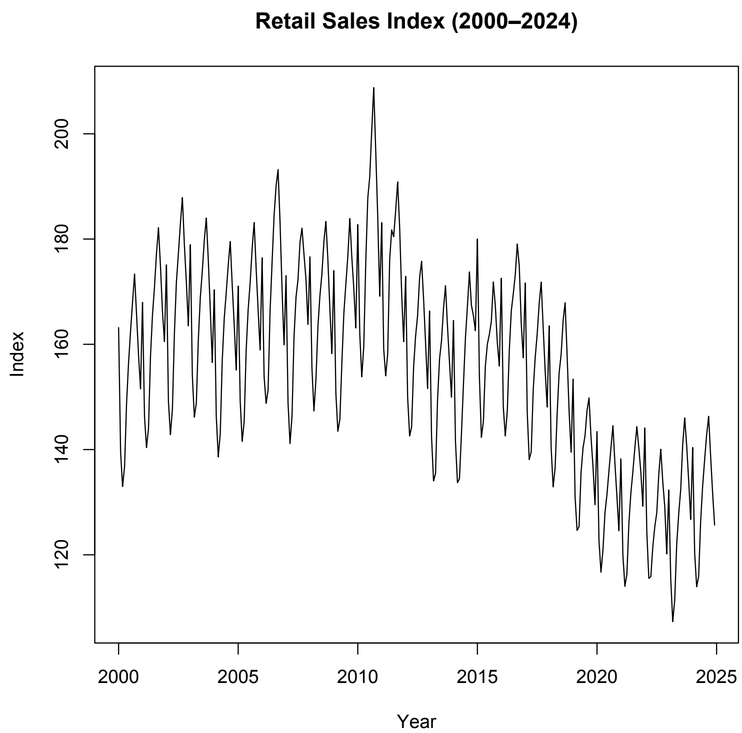


Figure 1: Time Series Plot of Retail Sales Index (2000–2024)

Based on the time plot, the Retail Sales Index shows a generally increasing trend from 2000 to 2010, reflecting a period of strong retail growth. Since 2010, the Retail Sales Index has shown a general downward trend. Although there are some slight upturns around 2015 and again near 2024, the index remains lower than its initial level in 2000.
In addition to the long-term trend, the time plot shows repeating seasonal patterns, with similar highs and lows occurring each year.
This indicates that the series has both trend and seasonal components, making it suitable for decomposition and seasonal adjustment.

3. Time Series Decomposition

Both additive and multiplicative decomposition models were applied to the Retail Sales Index to analyze the underlying components of the series, including trend, seasonal, and residual elements.

• Additive Decomposition

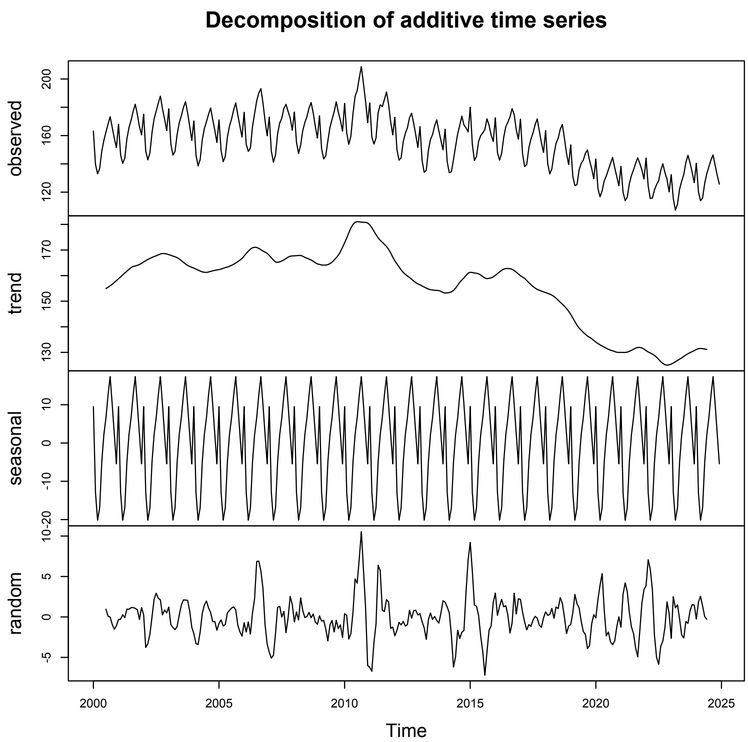


Figure 2: Additive Decomposition of the Retail Sales Index

From the additive decomposition, the trend component shows the general movement of the index over time, increasing until around 2010 and then declining. The seasonal component remains relatively constant in amplitude throughout the time period, showing repeated peaks and troughs at regular yearly intervals. This implies that seasonal fluctuations are assumed to be independent of the overall level of the series. The residuals display some irregular variations but are fairly balanced.

• Multiplicative Decomposition

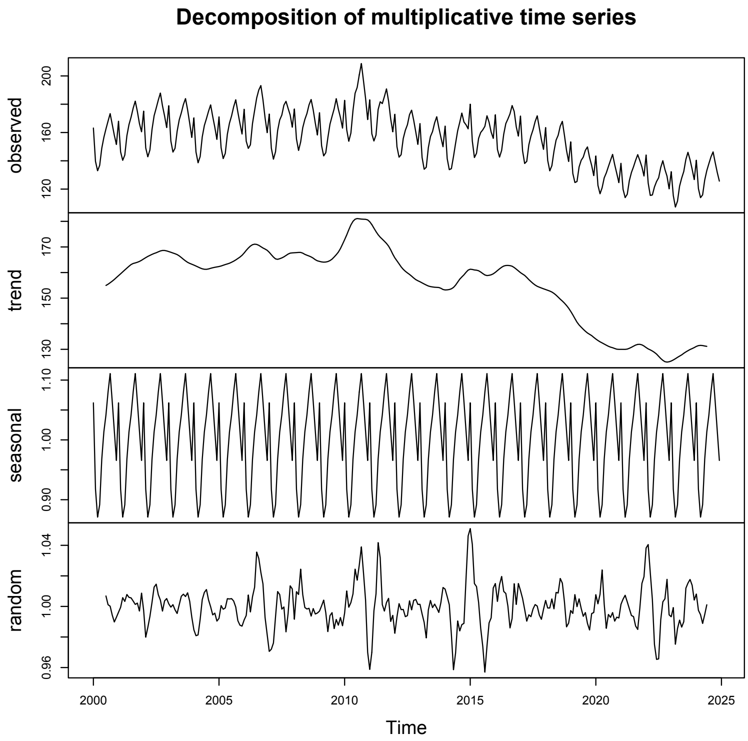


Figure 3: Multiplicative Decomposition of the Retail Sales Index

From the multiplicative decomposition, the trend component is quite similar to the additive decomposition. The seasonal component exhibits a repeating yearly pattern with consistent peaks and troughs. The magnitude of these seasonal effects remains relatively stable over time in the multiplicative model, although it is proportional to the level of the trend. The residuals from the multiplicative model appear more stable and randomly distributed.

The multiplicative model is preferred because the seasonal effect appears to change in magnitude as the overall trend changes — larger seasonal swings when the index is high and smaller swings when it is low. This relationship is better captured by a multiplicative decomposition.

Now, in order to observe the underlying trend and irregular components more clearly, we extract the seasonal effect from the multiplicative model and replot the time series as below.

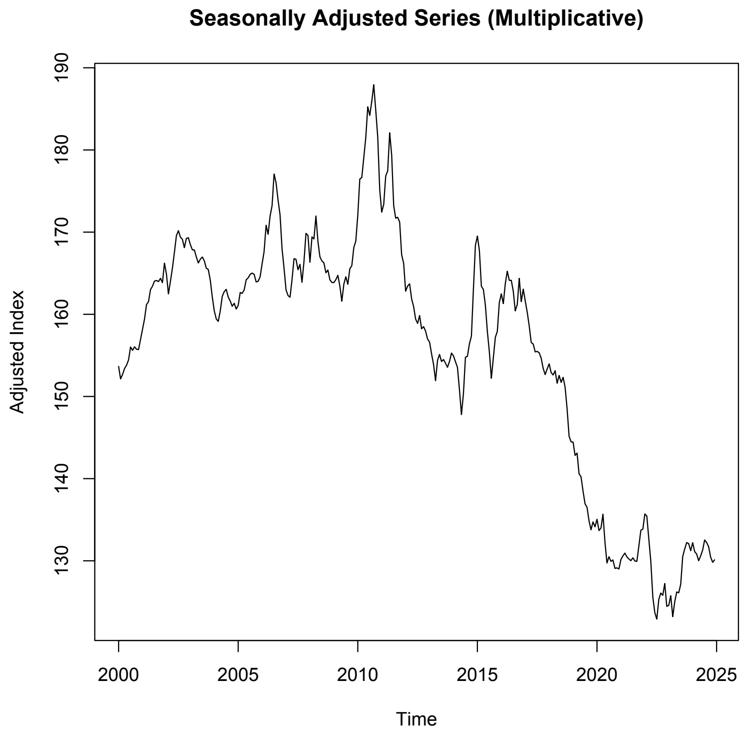
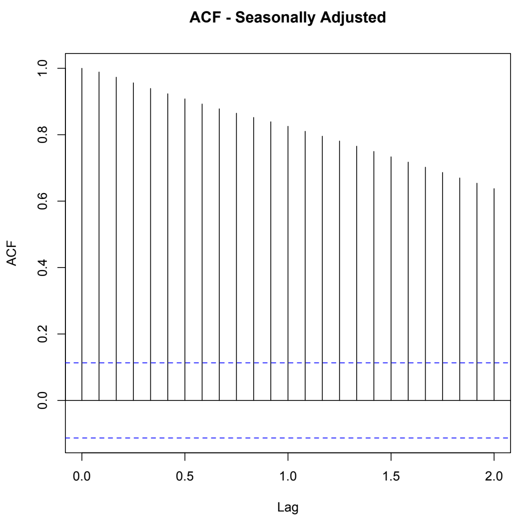


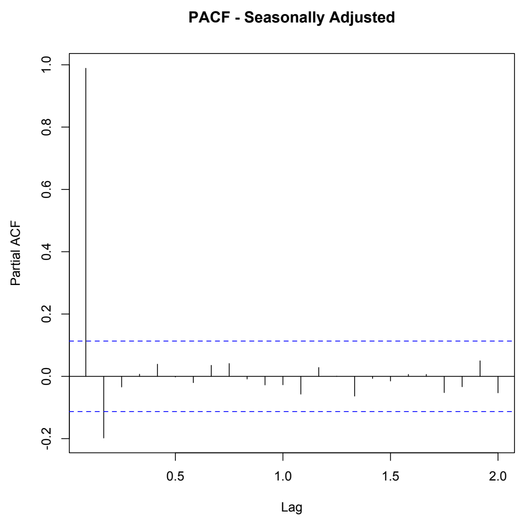
Figure 4: Time Series Plot of Seasonally Adjusted Retail Sales Index

4. Correlation Structure (ACF & PACF)

To further investigate the structure of the seasonally adjusted series, the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were examined.



(a) ACF Plot



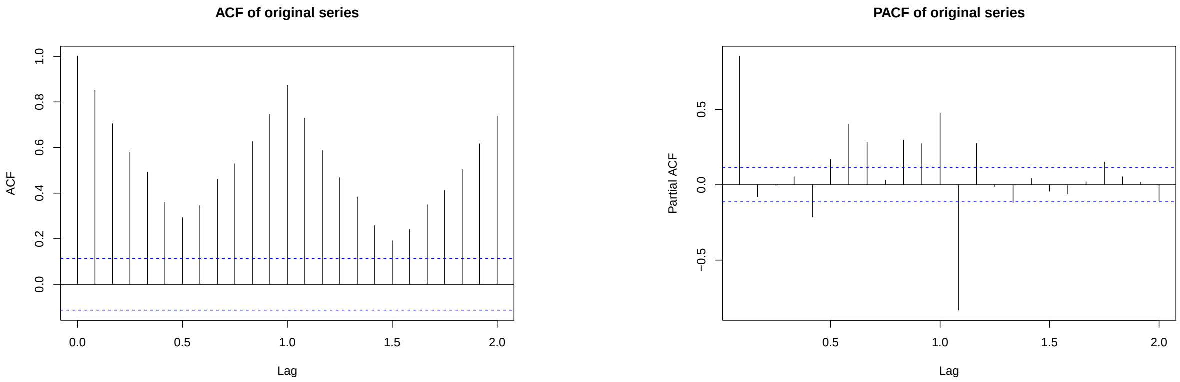
(b) PACF Plot

Figure 5: ACF and PACF of the Seasonally Adjusted Series

The ACF plot shows a gradual and slow decay, which is characteristic of an autoregressive (AR) process. In contrast, the PACF plot displays a significant spike at lag 0 and cuts off quickly afterward, with no notable correlations beyond the first lag. This pattern is consistent with an AR(1) model, where only the first lag of the series is strongly correlated and subsequent lags contribute minimally.

5. Stationarity and Differencing

We first plot the ACF and PACF for the original series.



(a) ACF of Original Series (b) PACF of Original Series

Figure 6: ACF and PACF of the original series

Based on ACF/PACF plots of the original series, the ACF plot exhibits a clear seasonal pattern of 12 months, and all values are above the significance bounds, indicating strong autocorrelation and the presence of seasonality. Additionally, the PACF plot shows many significant spikes exceeding the significance limits, suggesting the presence of multiple autoregressive components. Also, based on ACF/PACF plots of the Seasonally Adjusted Series, we observe that slow decay of the ACF plot and the PACF shows a significant spike at lag 1, exceeding the significance bounds. Above suggest that the series appears non-stationary.

There is two way for us to remove the seasonal, we applied both of then and compared the ACF and PACF of the time series after remove the seasonal and the first-order differencing as followed.

- Appling the first-order differencing and using lag-12 differencing to remove seasonal

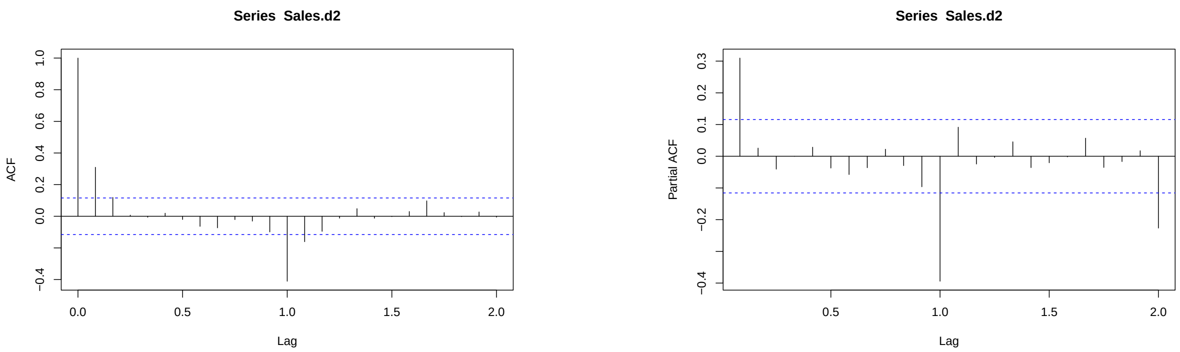
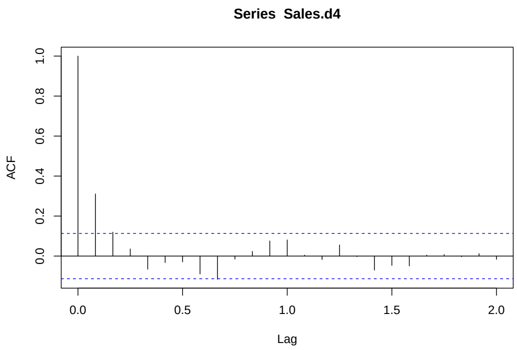
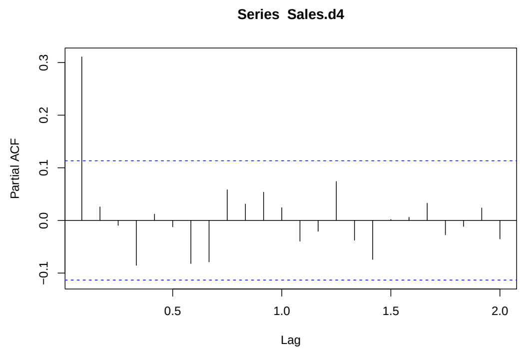


Figure 7: ACF and PACF of the original series

- Using decompose and manual remove the seasonal and appling the first-order differencing





(a) PACF of Original Series

We observed that the ACF and PACF perform better for the second method than the first one, so we decided to select the second mathod and below shows the original time series and the series after differencing and removed the seasonal.

- **Before differencing:**

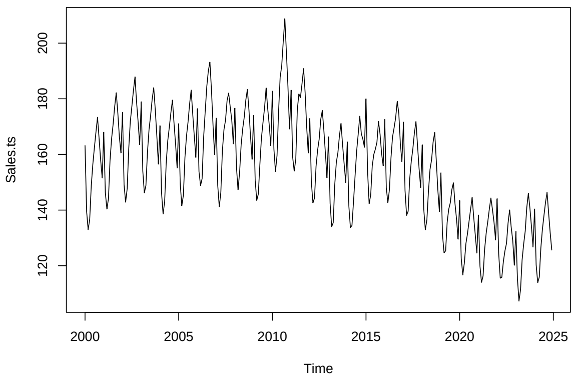


Figure 9: Original Time Series Plot

- **After first differencing:**

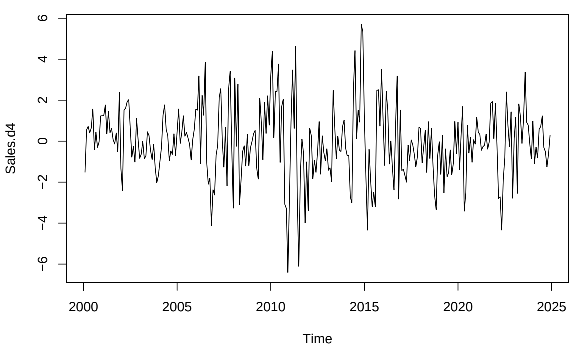


Figure 10: Time Series Plot After Differencing

6. Model Fitting and Comparison

In this section, we fit AR, MA, ARMA, and ARIMA models to the data separately. Then we will identify and justify the selection of the best-fitting model by compare the models using criteria such as AIC and residual diagnostics.

AR model

Based on the ACF and PACF plot of the seasonally adjusted and differentiated series, we observe the ACF decay gradually and a significant spike at lag 1, all the other lags falling within the significance bounds. This suggests that an AR(1) model is appropriate for the series.
From R we obtain:

```
> Sales.ar

Call:
arima(x = Sales.d4, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.3104    -0.0804
s.e.  0.0549     0.1409

sigma^2 estimated as 2.832:  log likelihood = -579.92,  aic = 1165.84
> AIC(Sales.ar)
[1] 1165.843
```

After removing the seasonal component and applying first-order differencing to the original series, the fitted AR(1) model is expressed as

$$y_t = -0.0804 + 0.3104\, y_{t-1} + w_t,$$

where

- y_t represents the differenced, seasonally adjusted series;
- w_t denotes the white noise term with variance $\sigma^2 = 2.832$;
- 0.3104 is the AR(1) coefficient, and -0.0804 is the intercept term.

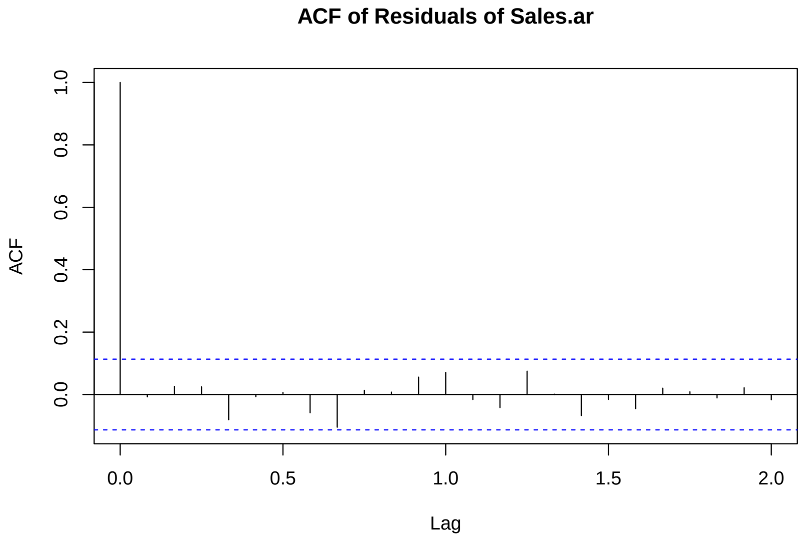
Let x_t denote the original series and s_t the seasonal component at time t. Since the adjusted series is obtained as

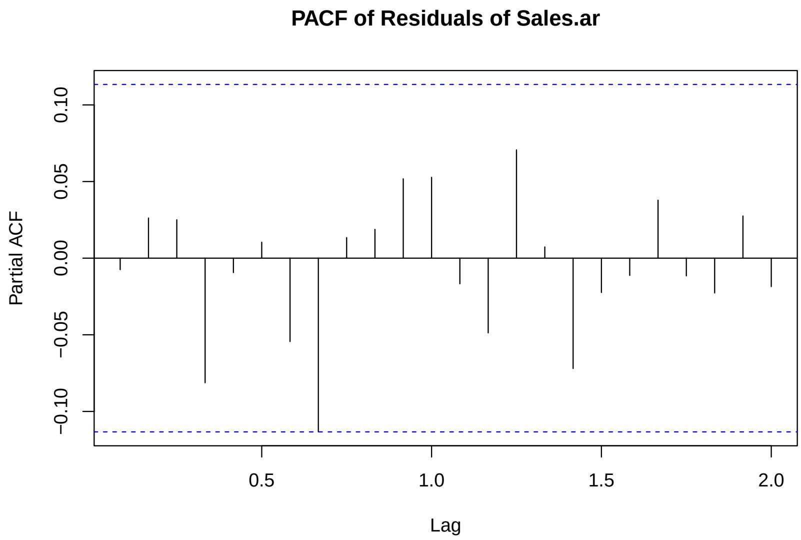
$$y_t = \frac{x_t}{s_t} - \frac{x_{t-1}}{s_{t-1}},$$

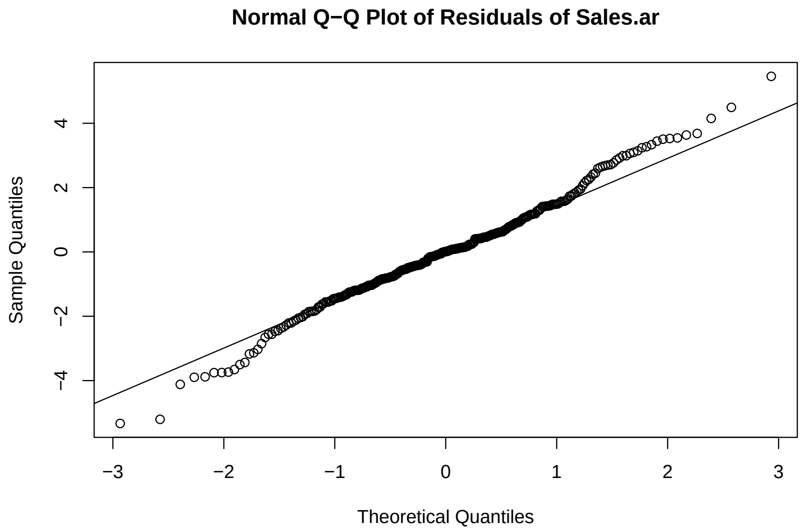
the restored model for the original series is given by

$$x_t = s_t \left[\frac{x_{t-1}}{s_{t-1}} - 0.0804 + 0.3104 \left(\frac{x_{t-1}}{s_{t-1}} - \frac{x_{t-2}}{s_{t-2}} \right) + w_t \right].$$

We also obtain the diagnostic plots for the residuals from the AR model, including the ACF, PACF, and normal Q-Q plot from R, which are presented separately below.


ACF of Residuals of Sales.ar
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
Lag


PACF of Residuals of Sales.ar
Partial ACF
0.10
0.05
0.00
−0.05
−0.10
0.5
1.0
1.5
2.0
Lag


Normal Q−Q Plot of Residuals of Sales.ar
Sample Quantiles
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

MA model

Since we did not find any suggestion from the above plot, we decided to set the maximum order to 4 and fit the MA models sequentially. The optimal order is then selected based on the fitting results. From R we obtain:

```
> Sales.ma

Call:
arima(x = Sales.d4, order = c(0, 0, 1))

Coefficients:
        ma1  intercept
      0.281    -0.0798
s.e.  0.052     0.1253

sigma^2 estimated as 2.865:  log likelihood = -581.67,  aic = 1169.34


> Sales.ma2

Call:
arima(x = Sales.d4, order = c(0, 0, 2))

Coefficients:
         ma1     ma2  intercept
      0.2924  0.1027    -0.0801
s.e.  0.0569  0.0599     0.1357

sigma^2 estimated as 2.837:  log likelihood = -580.21,  aic = 1168.43


> Sales.ma3

Call:
arima(x = Sales.d4, order = c(0, 0, 3))

Coefficients:
         ma1     ma2     ma3  intercept
      0.3096  0.1330  0.0794    -0.0806
s.e.  0.0584  0.0645  0.0569     0.1475

sigma^2 estimated as 2.819:  log likelihood = -579.25,  aic = 1168.51


> Sales.ma4

Call:
arima(x = Sales.d4, order = c(0, 0, 4))

Coefficients:
         ma1     ma2     ma3      ma4  intercept
      0.3019  0.1277  0.0605  -0.0724    -0.0802
s.e.  0.0578  0.0608  0.0593   0.0626     0.1372

sigma^2 estimated as 2.806:  log likelihood = -578.6,  aic = 1169.2
```

Order	**AIC**	σ^2
1	1169.34	2.865
2	1168.43	2.837
3	1168.51	2.819
4	1169.20	2.806

Table 1: Comparison Based on AIC and σ^2

So we select order of 2 to fit our MA model since it has the minimum AIC value.
After removing the seasonal component and applying first-order differencing to the original series, the fitted MA(2) model is expressed as

$$y_t = -0.0801 + w_t + 0.2924\, w_{t-1} + 0.1027\, w_{t-2},$$

where

- y_t represents the differenced, seasonally adjusted series;
- w_t denotes the white noise term with variance $\sigma^2 = 2.837$;
- the coefficients 0.2924 and 0.1027 are the MA(1) and MA(2) parameters, respectively.

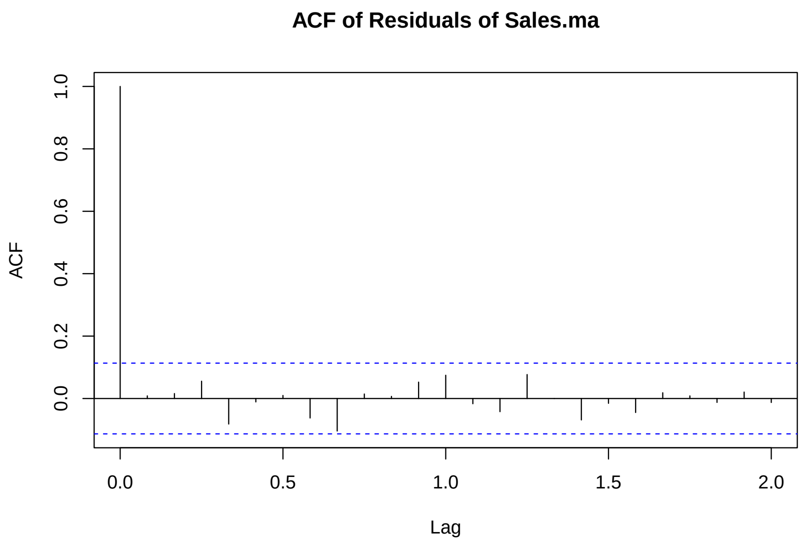
Let x_t denote the original series and s_t the seasonal component at time t. Since the adjusted series is obtained as

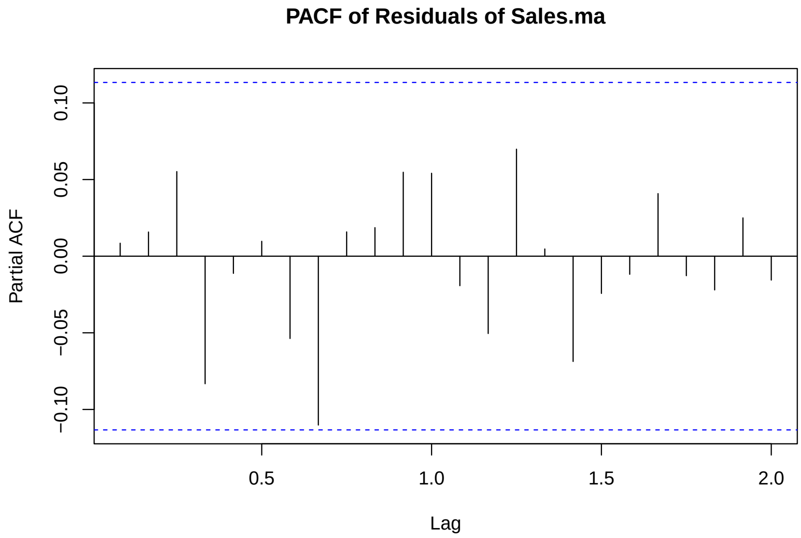
$$y_t = \frac{x_t}{s_t} - \frac{x_{t-1}}{s_{t-1}},$$

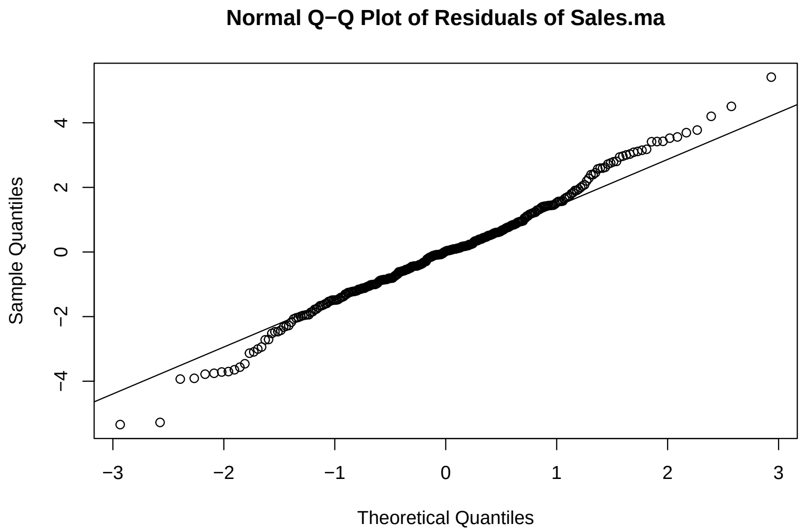
the restored model for the original series is given by

$$x_t = s_t \left[\frac{x_{t-1}}{s_{t-1}} - 0.0801 + w_t + 0.2924\, w_{t-1} + 0.1027\, w_{t-2} \right].$$

We also obtain the diagnostic plots for the residuals from the MA model, including the ACF, PACF, and normal Q-Q plot from R, which are presented separately below.







ARMA model

We first defined the bestmodel function to identify the most appropriate model parameters and then we obtained the most appropriate model parameters by using this function with the upper bounds on p and q set at 2.

The function and the most appropriate model parameters we obtained is shown below.

```
> bestmodel <- function(data,p,q){
+                                  best.order <- c(0,0,0)
+                                  best.aic <- Inf
+                                  for (i in 0:p) for (j in 0:q){
+                                    fit.aic <- AIC(arima(data, order = c(i,0,j)))
+                                    if (fit.aic < best.aic){
+                                      best.order <- c(i,0,j)
+                                      best.aic <- fit.aic
+                                    }
+                                  }
+                                 return(best.order)
+                                                          }
> bestmodel(Sales.d4,2,2)
[1] 1 0 0
```

Based on the selected model parameters, we fitted the ARMA model.

```
> Sales.arma

Call:
arima(x = Sales.d4, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.3104    -0.0804
s.e.  0.0549     0.1409

sigma^2 estimated as 2.832:  log likelihood = -579.92,  aic = 1165.84
```

After removing the seasonal component and applying first-order differencing to the original series, the fitted ARMA(1,0) model is expressed as

$$y_t = -0.0804 + 0.3104\, y_{t-1} + w_t, \tag{1}$$

where

- y_t represents the differenced, seasonally adjusted series;
- w_t denotes the white noise term with variance $\sigma^2 = 2.832$;
- 0.3104 is the AR(1) coefficient, and -0.0804 is the intercept term.

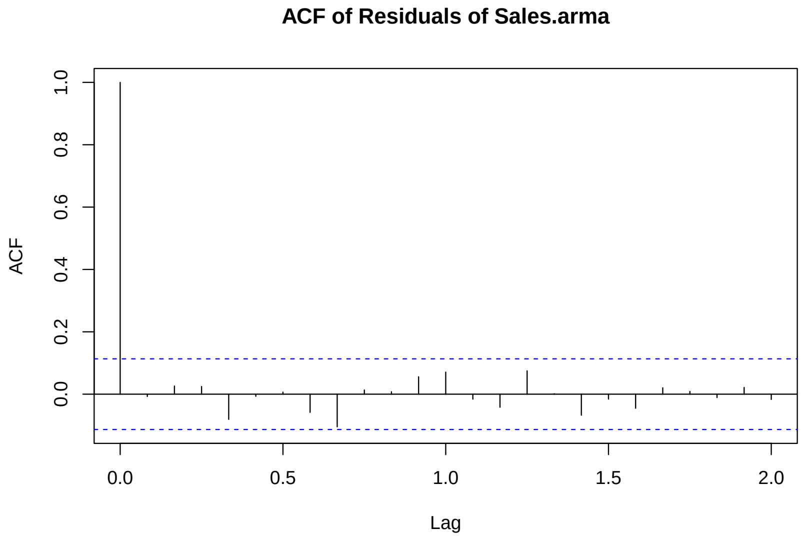
Let x_t denote the original series and s_t the seasonal component at time t. Since the adjusted series is obtained as

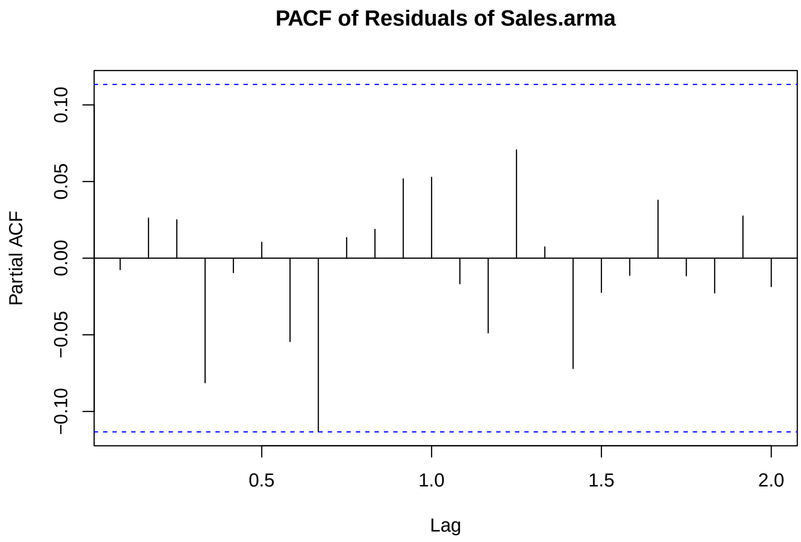
$$y_t = \frac{x_t}{s_t} - \frac{x_{t-1}}{s_{t-1}}, \tag{2}$$

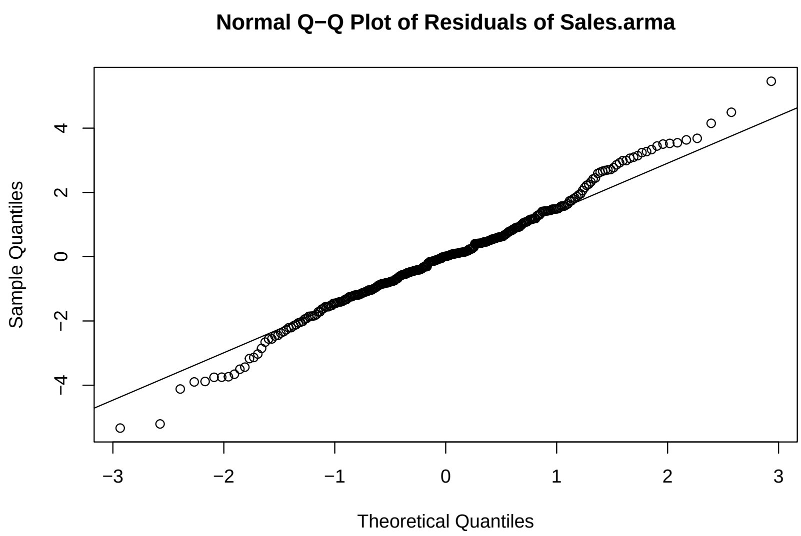
the restored model for the original series is given by

$$\begin{aligned} x_t = s_t \Bigg[& \frac{x_{t-1}}{s_{t-1}} - 0.0804 \\ & + 0.3104 \left(\frac{x_{t-1}}{s_{t-1}} - \frac{x_{t-2}}{s_{t-2}} \right) + w_t \Bigg]. \end{aligned} \tag{3}$$

We also obtain the diagnostic plots for the residuals from the ARMA model, including the ACF, PACF, and normal Q-Q plot from R, which are presented separately below.


ACF of Residuals of Sales.arma
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
Lag


PACF of Residuals of Sales.arma
Partial ACF
0.10
0.05
0.00
−0.05
−0.10
0.5
1.0
1.5
2.0
Lag


Normal Q−Q Plot of Residuals of Sales.arma
Sample Quantiles
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

ARIMA model

For the ARIMA model, we fit the models on the data after manually removing the seasonal component but without performing differencing. Since we have performed a first-order differencing on the original series in the previous analysis, we set $d = 1$ as the parameter for differencing. For p, based on our earlier fitting of AR models, the AR(1) model was found to be appropriate. Moreover, when fitting ARMA models, the AR parameter was also determined to be $p = 1$. Therefore, we fix $p = 1$. For q, our analysis of MA models suggested that MA(2) was suitable, but in the ARMA model analysis, q was determined to be $q = 0$. To account for these observations, we consider $q = 0, 1, 2$ and fit the corresponding models sequentially. In summary, we will fit the ARIMA models with orders $(1, 1, 0)$, $(1, 1, 1)$, and $(1, 1, 2)$, and compare their performance to select the most appropriate model.

```
> Sales.arima1

Call:
arima(x = Sales.ns, order = c(1, 1, 0))

Coefficients:
         ar1
      0.3117
s.e.  0.0549

sigma^2 estimated as 2.835:  log likelihood = -580.08,  aic = 1164.17
```

```
> Sales.arima2

Call:
arima(x = Sales.ns, order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.3782  -0.0737
s.e.  0.1581   0.1691

sigma^2 estimated as 2.833:  log likelihood = -579.99,  aic = 1165.99
```

```
> Sales.arima3

Call:
arima(x = Sales.ns, order = c(1, 1, 2))

Coefficients:
         ar1      ma1     ma2
      0.3062  -0.0041  0.0335
s.e.  0.2964   0.2949  0.1096

sigma^2 estimated as 2.832:  log likelihood = -579.95,  aic = 1167.9
```

Order	**AIC**	σ^2
(1,1,0)	1164.17	2.835
(1,1,1)	1165.99	2.833
(1,1,2)	1167.90	2.832

Table 2: Comparison Based on AIC and σ^2

So we select order of (1,1,0) to fit our ARIMA model since it has the minimum AIC value.

After removing the seasonal component from the original series, the fitted ARIMA(1,1,0) model is expressed as

$$\Delta y_t = 0.3117\, \Delta y_{t-1} + w_t, \tag{4}$$

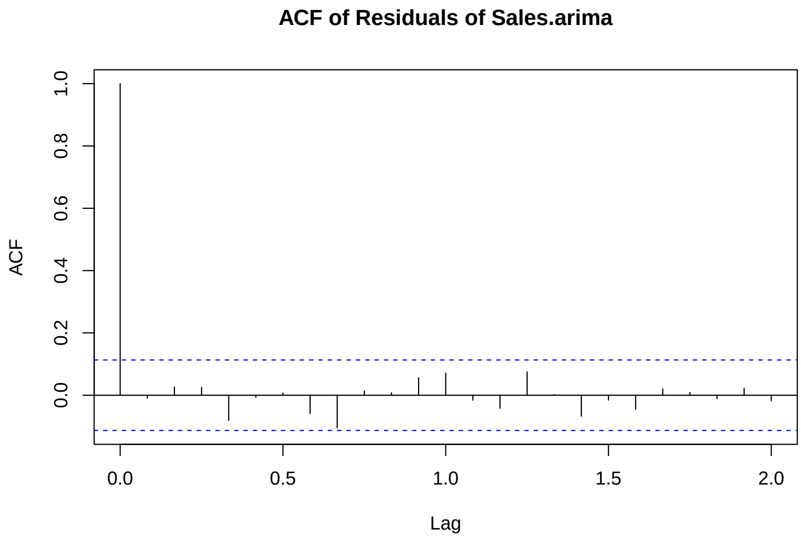
where

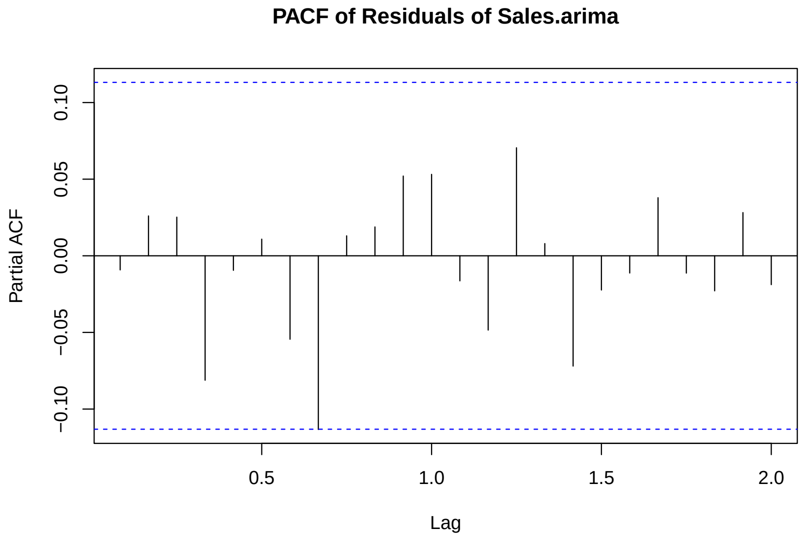
- $y_t = \dfrac{x_t}{s_t}$ represents the seasonally adjusted series;
- $\Delta y_t = y_t - y_{t-1}$ denotes the first-order differenced, seasonally adjusted series;
- w_t denotes the white noise term with variance $\sigma^2 = 2.835$;
- 0.3117 is the AR(1) coefficient.

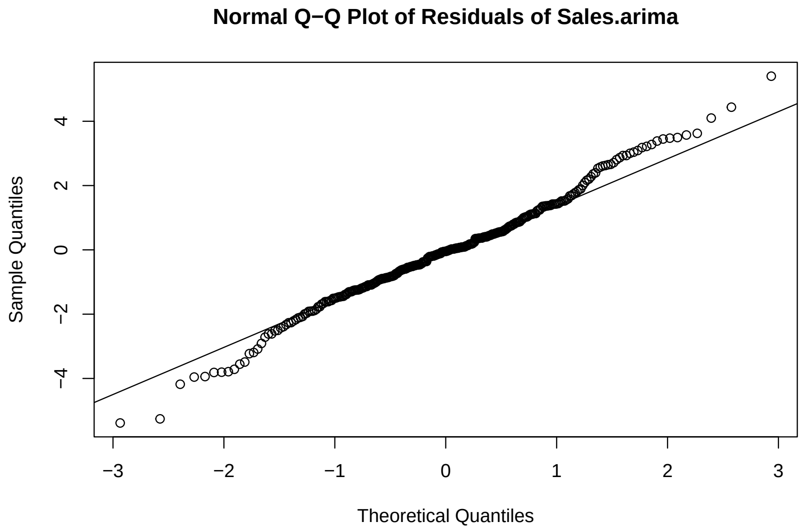
Let x_t denote the original series and s_t the seasonal component at time t. The model for the original series can be written as

$$\frac{x_t}{s_t} = \frac{x_{t-1}}{s_{t-1}} + 0.3117 \left(\frac{x_{t-1}}{s_{t-1}} - \frac{x_{t-2}}{s_{t-2}} \right) + w_t. \tag{5}$$

We also obtain the diagnostic plots for the residuals from the ARIMA model, including the ACF, PACF, and normal Q-Q plot from R, which are presented separately below.







Model Comparison

Model	AIC	σ^2
AR	1165.84	2.832
MA	1168.43	2.837
ARMA	1165.84	2.832
ARIMA	1164.17	2.835

Table 3: Model Comparison Based on AIC and σ^2

Based on the AIC values and residual variance, the ARIMA model currently appears to be the best among the fitted models. To further validate this conclusion, we proceed to examine the diagnostic plots of the residuals for each model.

We observed that, for all models, the points on the Q-Q plots of the residuals lie approximately along the reference line, suggesting that the residuals are roughly normally distributed. Moreover, when examining the ACF and PACF plots of the residuals, all models show that the autocorrelations fall within the confidence bounds, indicating that the residuals can be regarded as white noise. This suggests that each model provides an adequate fit from the perspective of residual diagnostics.

Nevertheless, since the ARIMA model achieves the lowest AIC value, it is considered the best model overall.

Best Model: ARIMA Model.

7. Forecasting (Next 12 Months)

Using the ARIMA model, the next 12 months are forecasted with 95% confidence intervals.

```
> pred.f
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2025 138.2672 119.5096 113.4276 116.1237 126.0212 131.9827 135.6664 140.6476
          Sep      Oct      Nov      Dec
2025 144.6848 138.5936 132.1700 125.7676
> upper
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2025 141.5672 124.9526 120.5859 124.7141 135.8520 142.9184 147.6063 153.5140
          Sep      Oct      Nov      Dec
2025 158.4154 153.1371 147.4833 141.8140
> lower
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2025 134.9673 114.0666 106.2693 107.5333 116.1903 121.0470 123.7264 127.7812
          Sep      Oct      Nov      Dec
2025 130.9543 124.0501 116.8566 109.7213
```

Figure 11: 12 Months Forecasts

The object `pred.f` contains the predicted values, while `upper` and `lower` represent the upper and lower bounds of the 95% confidence interval for the predictions, respectively. The plot below illustrates these values.

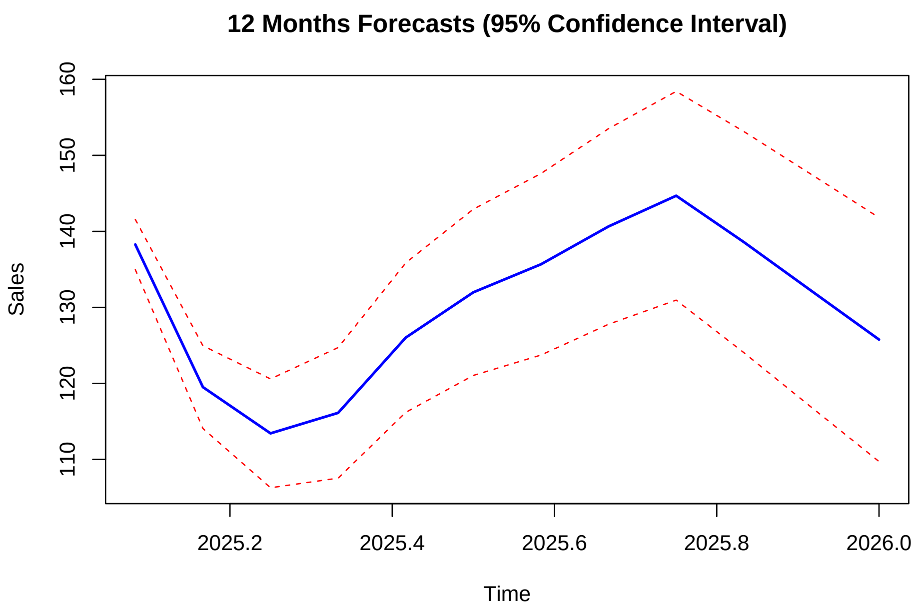


Figure 12: 12 Months Forecasts

Based on the forecast plot above, we observe that the predicted values for the next 12 months initially decrease, then increase slightly, and decrease again. This reflects a cyclical pattern similar to that observed in the original time series, indicating that the model has effectively captured the underlying seasonality. The 95% confidence intervals are relatively narrow, suggesting stable and reliable forecasts, and further supporting the good fit of the ARIMA model.

8. Conclusion

This report systematically analyzes the monthly Retail Sales Index from January 2000 to December 2024 using a variety of time series modeling techniques. We began by visualizing and decomposing the series to identify long-term trends and seasonal patterns. This motivated the use of seasonal adjustment and differencing techniques to achieve stationarity, enabling effective model fitting.

Several candidate models—including AR, MA, ARMA, and ARIMA—were fitted and evaluated. Model comparisons were conducted using AIC values and residual diagnostics. The residual diagnostics, including ACF, PACF, and Q-Q plots, suggested that all models provided an adequate fit, with no significant autocorrelations remaining in the residuals.

Among these, the ARIMA model achieved the lowest AIC value, indicating the best balance between model fit and complexity. Finally, we used the ARIMA model to generate forecasts for the next 12 months, along with confidence intervals. The forecasts were visualized and shown to effectively capture the underlying seasonality of the original series.

9. Work Distribution

Student ID	**Member**	**Contribution**
EGE2209334	Lu Zhimin	Introduction, Problem 1,2,3 and assist on Problem 4,5,6
MAT2309449	Wang Fangyu	Major part of Problem 4,5,6 and Conclusion.