DataAnalysis2

Enyu Li

2025-04-23

```
knitr::opts_chunk$set(echo = FALSE)
library(astsa)
library(xts)  # This will automatically load zoo
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
#library(ggfortify)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::first()  masks xts::first()
## x dplyr::lag()    masks stats::lag()
## x dplyr::last()   masks xts::last()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(knitr)
library(tidyquant)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
## -- Attaching core tidyquant packages ----------------------- tidyquant 1.0.10 --
## v PerformanceAnalytics 2.0.8      v TTR                  0.24.4
## v quantmod             0.4.26     -- Conflicts ------------------------------------------ tidyquant_
## x zoo::as.Date()                 masks base::as.Date()
## x zoo::as.Date.numeric()         masks base::as.Date.numeric()
## x dplyr::filter()                masks stats::filter()
## x dplyr::first()                 masks xts::first()
## x dplyr::lag()                   masks stats::lag()
## x dplyr::last()                  masks xts::last()
## x PerformanceAnalytics::legend() masks graphics::legend()
## x quantmod::summary()            masks base::summary()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(fanplot)
library(urca)
library(forecast)
```

```
##
## Attaching package: 'forecast'
##
## The following object is masked from 'package:astsa':
##
##     gas
```

```
library(fpp3)
```

```
## Registered S3 method overwritten by 'tsibble':
##   method               from
##   as_tibble.grouped_df dplyr
## -- Attaching packages -------------------------------------------- fpp3 1.0.1 --
## v tsibble     1.1.5     v feasts      0.4.1
## v tsibbledata 0.4.1     v fable       0.4.1
## -- Conflicts ------------------------------------------------- fpp3_conflicts --
## x lubridate::date()    masks base::date()
## x dplyr::filter()      masks stats::filter()
## x dplyr::first()       masks xts::first()
## x tsibble::index()     masks zoo::index()
## x tsibble::intersect() masks base::intersect()
## x tsibble::interval()  masks lubridate::interval()
## x dplyr::lag()         masks stats::lag()
## x dplyr::last()        masks xts::last()
## x tsibble::setdiff()   masks base::setdiff()
## x tsibble::union()     masks base::union()
## x fable::VAR()         masks tidyquant::VAR()
##
## Attaching package: 'fpp3'
##
## The following object is masked from 'package:PerformanceAnalytics':
##
##     prices
```

Data Analysis 2 is worth 10% of final grade.

Honor Code: You may work with ONE other person on this analysis. If you do so, you should indicate both authors on the paper, but submit individually. Feel free to make use of generative AI tools, and online searches, as helpful.

DUE DATE: April 25 at midnight. Extensions will be granted through Tuesday April 30 at midnight.

WHAT TO SUBMIT: Solutions should be written using RMarkdown or Quarto. You will submit

- The compiled html (or pdf if you prefer)
- The RMarkdown or Quarto file

Modeling and Forecast ICE CREAM MANUFACTURING

Background: The time series used is IPN31152N.csv. This time series represents monthly ice cream production for the US since 1972 through 2024. The series is not seasonally adjusted but is indexed to 2017.

Data Citation: Board of Governors of the Federal Reserve System (US), Industrial Production: Manufacturing: Non-Durable Goods: Ice Cream and Frozen Dessert (NAICS = 31152) [IPN31152N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/IPN31152N, April 22, 2025.

Our goals for analyzing this data are

```
+ Describing and modeling to understand the dynamics of ice cream production
+ Forecasting monthly production through December of 2025
```

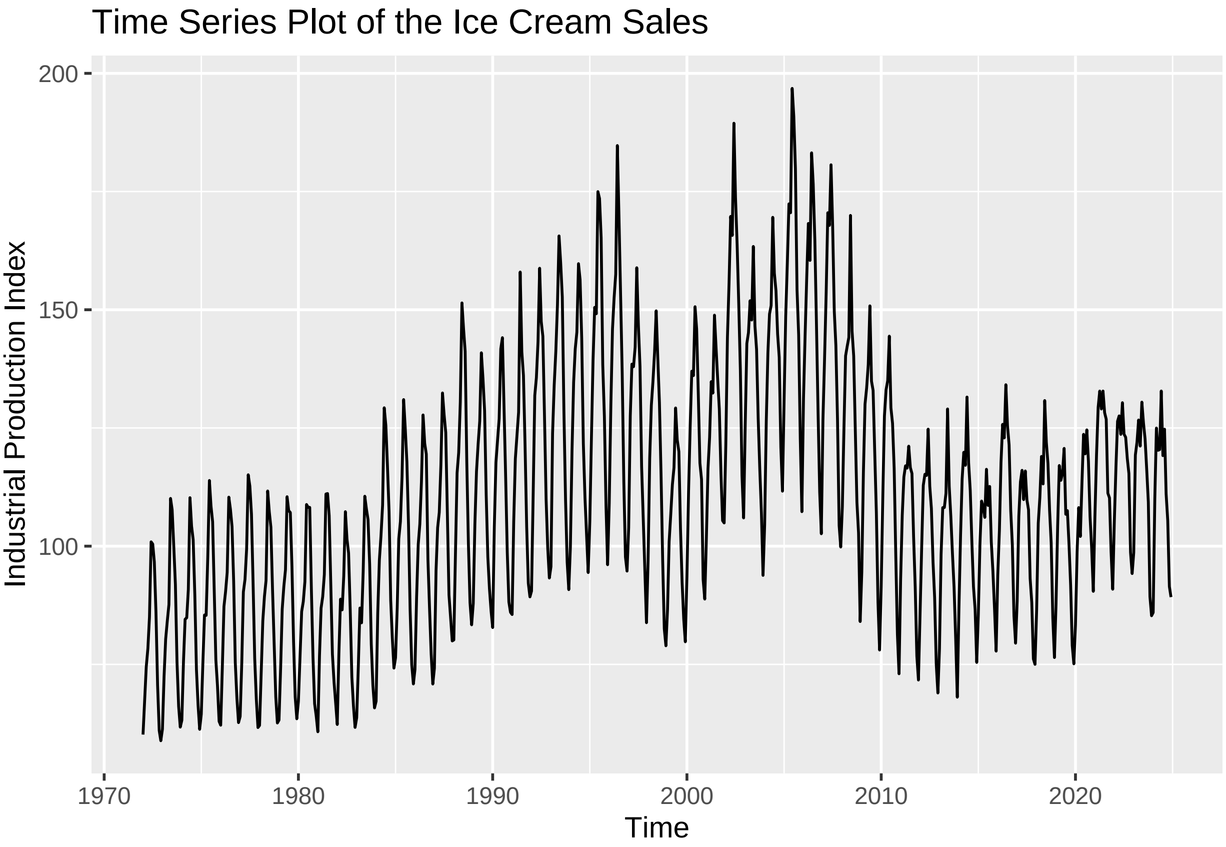
QUESTION 0 AFTER you answer questions 1-4, come back and ANSWER THIS QUESTION. Provide a succinct summary of your findings to address the above goals. Choose as the audience for your paragraph a manager at a company producing ice cream, or an investment manager who focuses on ice cream, or something similar. For example, these managers WILL NOT CARE ABOUT order selection criteria but will expect that you did your analysis well and the insight you are providing them is something that they can move forward with profitable decisions for the company. Do include uncertainty bounds in your narrative.

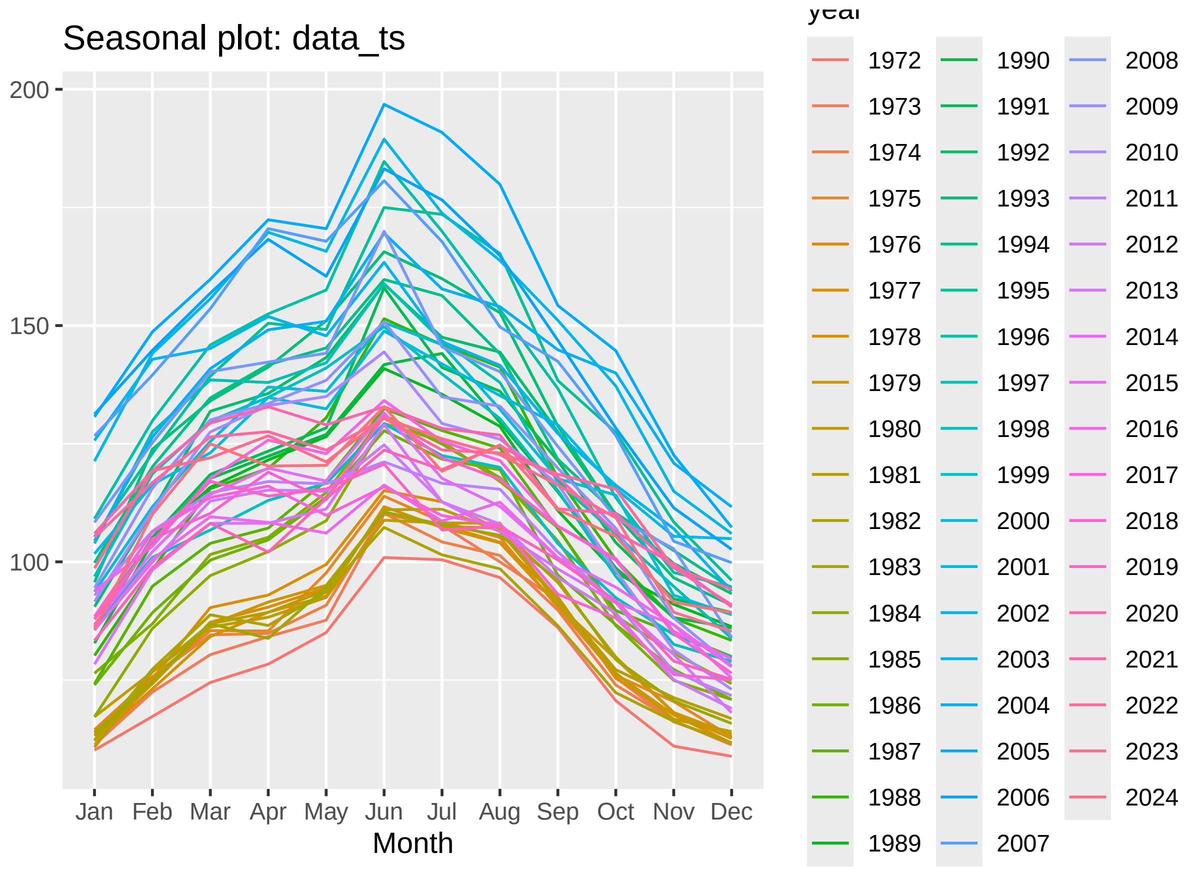
Based on a comprehensive time series analysis of historical U.S. ice cream production data, we identified strong seasonal patterns with peaks typically occurring in the warmer months. Using a statistically validated ARIMA(1,0,1)(0,1,1)[12] model, we forecast monthly ice cream production through December 2025. The model projects a continuation of established seasonal trends, with production expected to rise during summer months and dip during winter, consistent with historical demand cycles. While the central forecast provides actionable insight for inventory planning and marketing timing, the 80% and 95% confidence intervals around each monthly estimate indicate increasing uncertainty further into the future. For example, forecasted summer 2025 production volumes fall within a reasonably narrow band, while year-end estimates carry wider margins. This suggests the model is reliable in the short-to-medium term, and strategic decisions such as production scaling, distribution planning, and investment timing can be made with confidence—while remaining mindful of forecast variability, especially beyond a one-year horizon.

QUESTION 1 Produce descriptive plots and **{DISCUSS} what information you glean from each plot.**

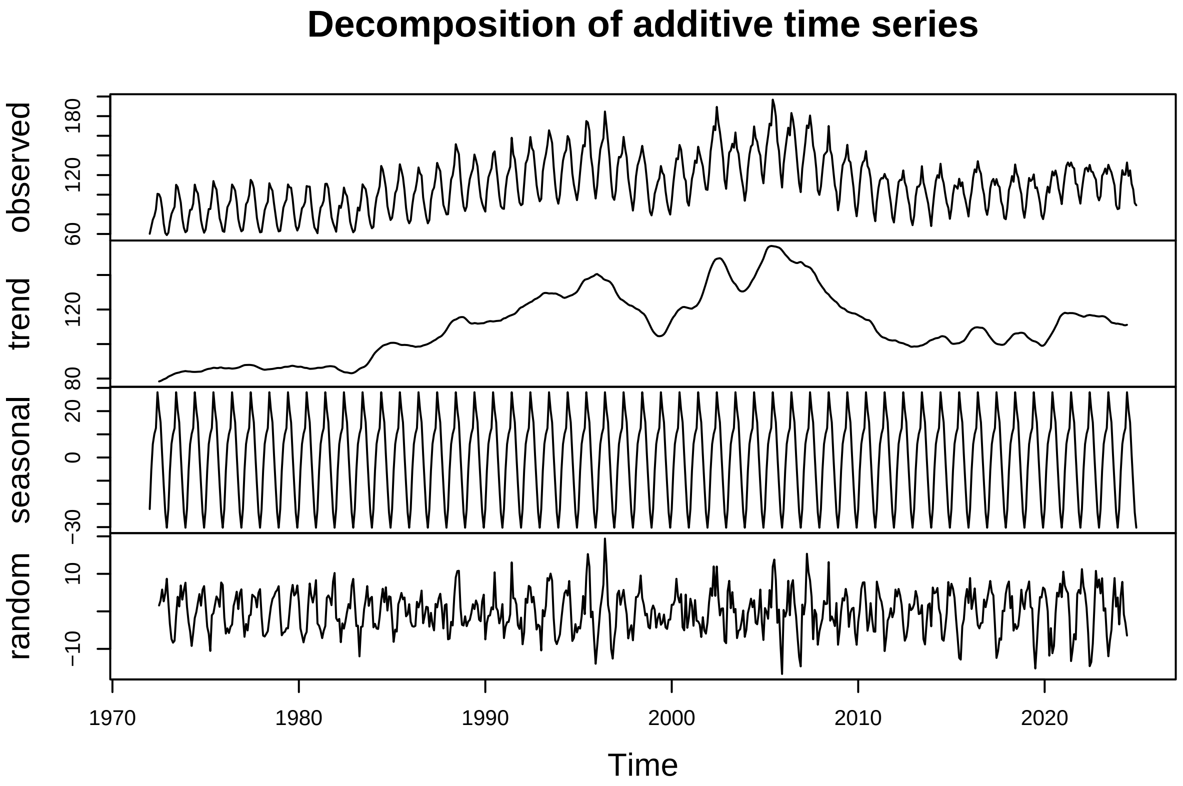
```
+ Time series
+ Relevant seasonal plot
+ Decomposition plot
+ ACF, PACF and periodogram
```

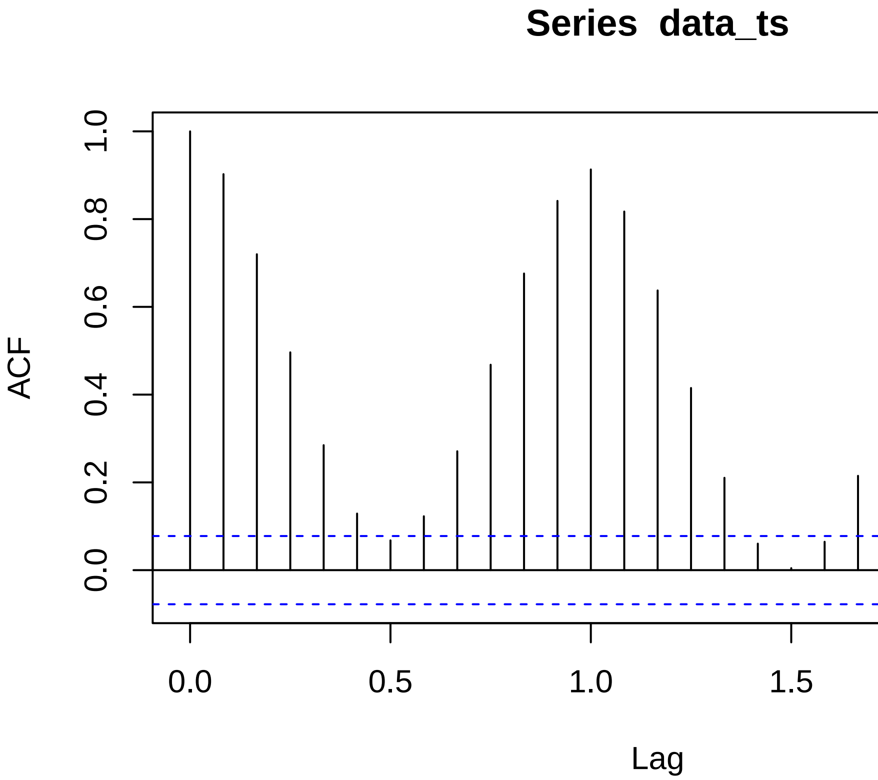
```
##   observation_date IPN31152N
## 1       1972-01-01   60.1519
## 2       1972-02-01   67.2727
## 3       1972-03-01   74.4700
## 4       1972-04-01   78.3594
## 5       1972-05-01   85.0321
## 6       1972-06-01  100.9147
```


Time Series Plot of the Ice Cream Sales
Industrial Production Index
200
150
100
1970
1980
1990
2000
2010
2020
Time

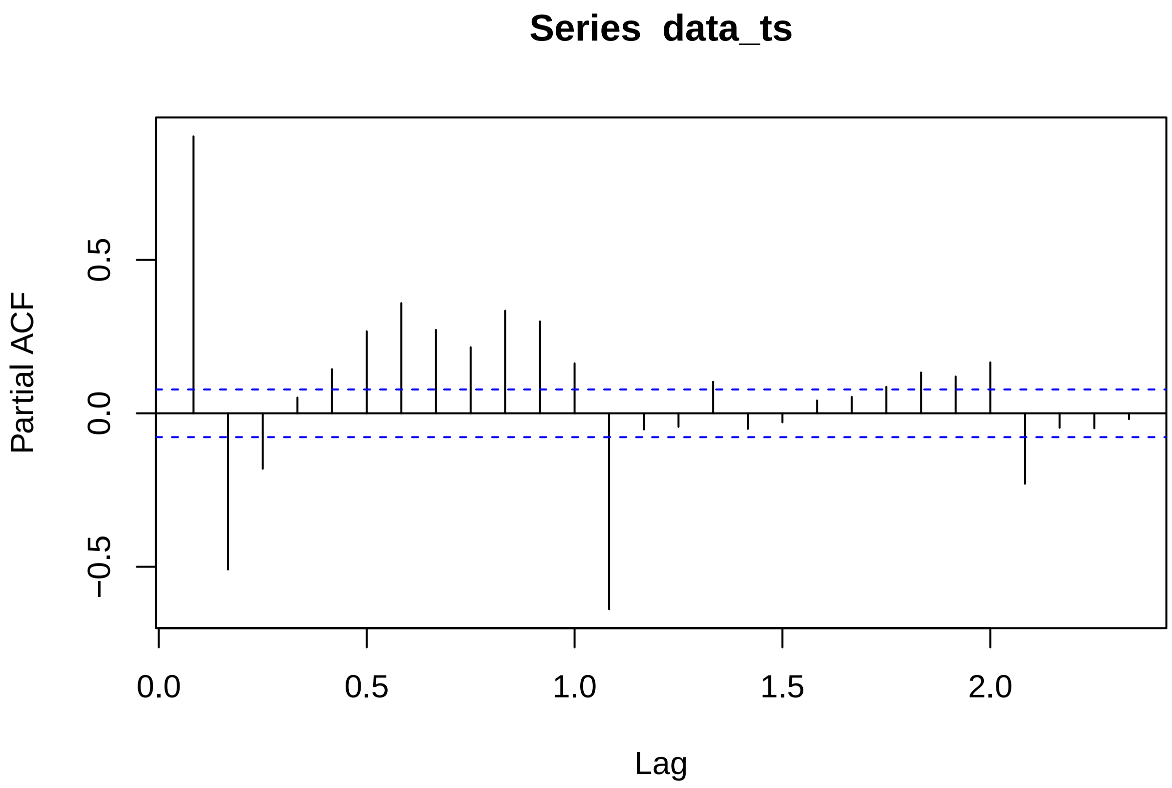


The seasonal plots clearly show that the sales of ice-cream sales clearly increase from winter in January to the summer in June. And this trend works for all of the years.

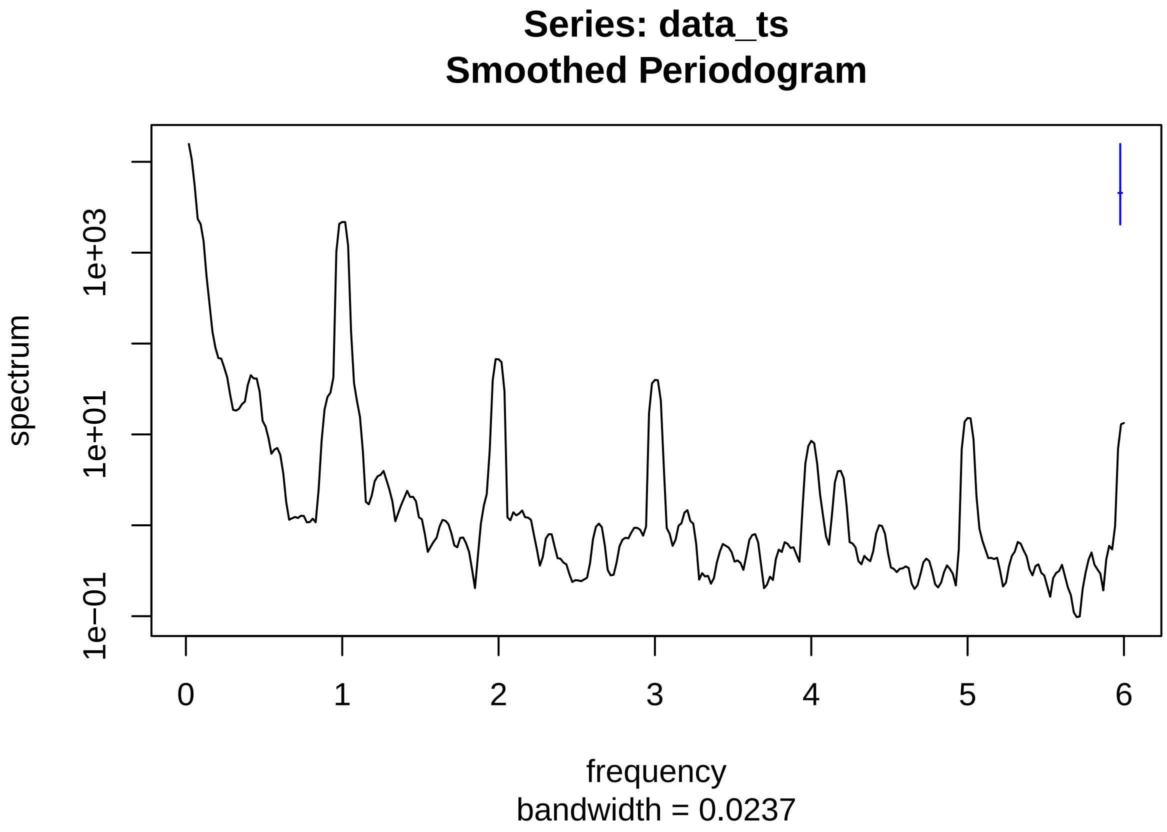

Decomposition of additive time series
observed
60
120
180
trend
80
120
seasonal
−30
0
20
random
−10
10
1970
1980
1990
2000
2010
2020
Time



The decomposition plot looks good to me.



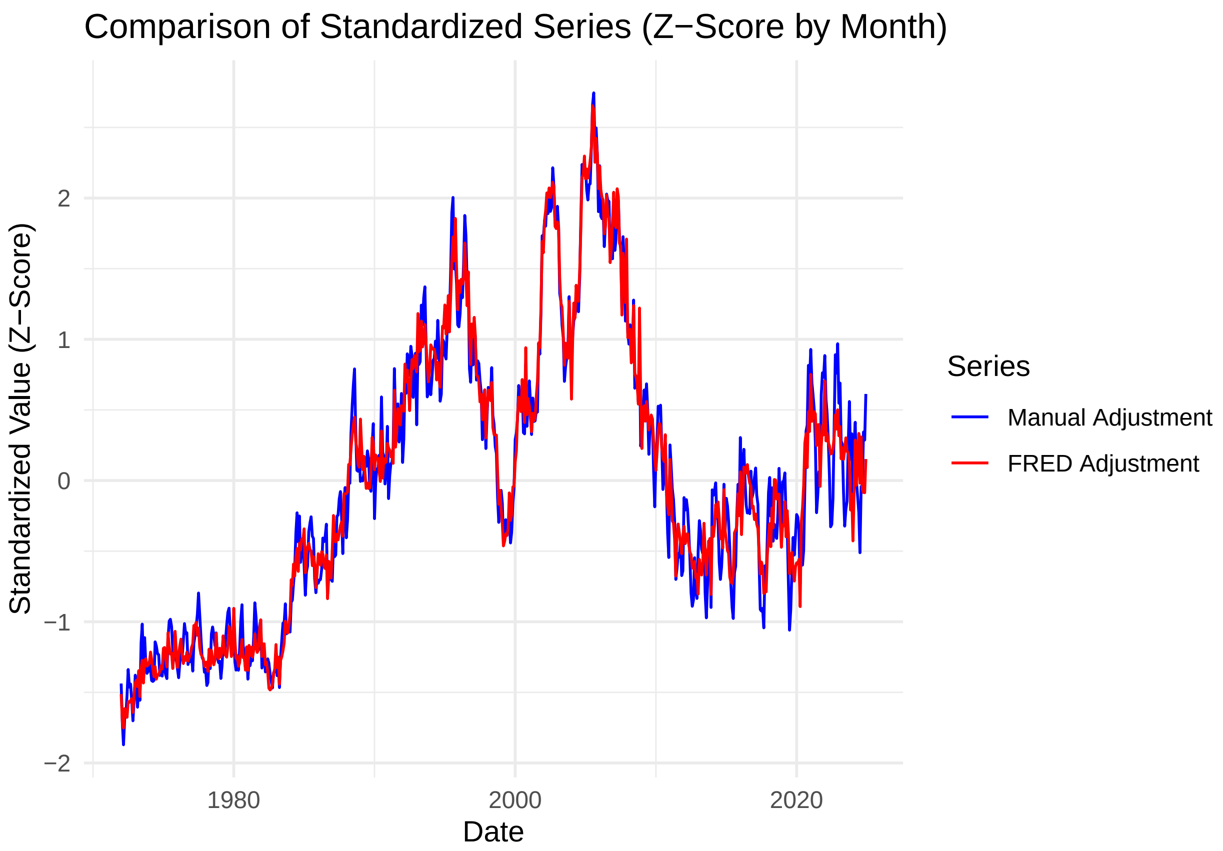
```
## Warning in plot.window(...): "dmean" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "dmean" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "dmean" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "dmean" is not a
## graphical parameter

## Warning in box(...): "dmean" is not a graphical parameter

## Warning in title(...): "dmean" is not a graphical parameter
```



```
## [1] 53
```

QUESTION 2 Construct a seasonally adjusted time series by standardizing each month with the mean and standard deviation for that month across all years, and compare to the seasonally adjusted series from FRED, also in the folder as IPN31152S.csv. Use whatever graphs you deem helpful for this comparison (sometimes something simple like a scatterplot with colors for years or months works well - your choice). Again be sure and DISCUSS.

```
##   observation_date IPN31152N       Date Year Month
## 1       1972-01-01   60.1519 1972-01-01 1972     1
## 2       1972-02-01   67.2727 1972-02-01 1972     2
## 3       1972-03-01   74.4700 1972-03-01 1972     3
## 4       1972-04-01   78.3594 1972-04-01 1972     4
## 5       1972-05-01   85.0321 1972-05-01 1972     5
## 6       1972-06-01  100.9147 1972-06-01 1972     6
```



To examine the effectiveness of seasonal adjustment methods, I manually standardized the original time series (IPN31152N) by calculating z-scores for each calendar month across all years, and compared the result to the officially seasonally adjusted series from FRED (IPN31152S). The plot shows that both series follow a nearly identical pattern over time, with peaks, troughs, and overall trends aligning closely. This indicates that the manual month-wise standardization successfully removes seasonal effects and captures the underlying structure of the data. While the manually adjusted series appears slightly more volatile due to the nature of z-score scaling, it provides a reasonable approximation of the FRED-adjusted values. This comparison demonstrates that simple standardization by month can be a useful and interpretable method for seasonal adjustment when more advanced tools are not available.

QUESTION 3 Develop an appropriate SARIMA model, or trend stationary SARMA model for the original unadjusted series. Justify your model choice using

```
+ Preliminary tests (e.g. unit roots, trend stationary)
+ Model selection criteria (eg. AIC, AICc, BIC)
+ Diagnostics of standaridized residuals
```

Be sure and include a discussion not just the plots.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts
## Dickey-Fuller = -2.3424, Lag order = 8, p-value = 0.4334
## alternative hypothesis: stationary
```

As we can see, we fail to reject the null hypothesis that the time series is non-stationary. Thus, the time series is non-stationary.

```
## Warning in adf.test(data_ts_diff): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts_diff
## Dickey-Fuller = -23.695, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

As we can see, the differenced time series is stationary.

Now check if seasonal differencing is needed.

```
## [1] 1
```

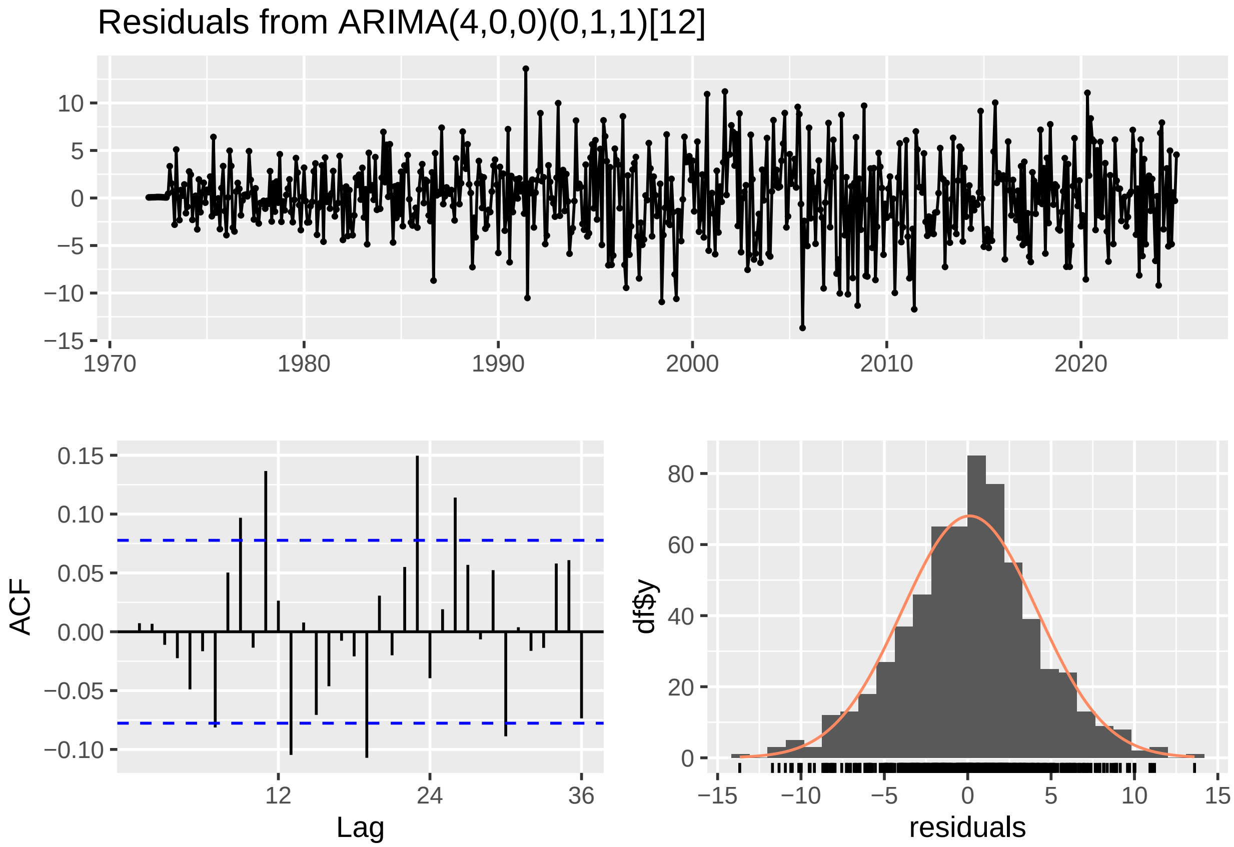
Yes, a seasonal differencing is needed.

This is what the auto SARIMA model

```
## Series: data_ts
## ARIMA(4,0,0)(0,1,1)[12]
##
## Coefficients:
##          ar1     ar2     ar3      ar4     sma1
##       0.7515  0.1596  0.2101  -0.1788  -0.6851
## s.e.  0.0399  0.0489  0.0490   0.0394   0.0311
##
## sigma^2 = 16.99:  log likelihood = -1771.24
## AIC=3554.47   AICc=3554.61   BIC=3581.09
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 0.1191165 4.066619 3.135784 0.0909375 2.833151 0.4305214
##                     ACF1
## Training set 0.007275947
```

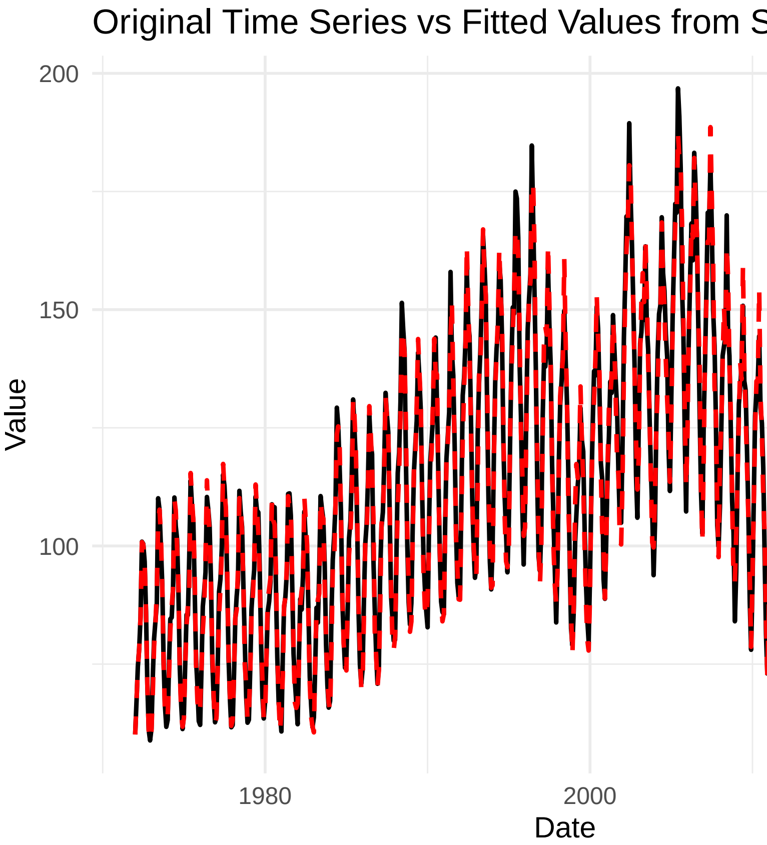
By using the auto.fit, we can see the best model has AIC = 3554.47, AICc = 3554.61, and BIC = 3581.09.

Now do a diagnostics of the residuals.



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(4,0,0)(0,1,1)[12]
## Q* = 65.287, df = 19, p-value = 5.483e-07
##
## Model df: 5.   Total lags used: 24
```

The residual diagnostics from the SARIMA(4,0,0)(0,1,1)[12] model suggest that the model provides an overall adequate fit to the data. The residuals fluctuate around zero without visible trends or seasonal patterns, indicating that the model has captured the main structure of the series. The ACF plot shows most autocorrelations are within the 95% confidence bounds, suggesting that the residuals are largely uncorrelated, though small spikes at seasonal lags (e.g., lag 12 and 24) hint at minor remaining seasonal structure. The histogram of residuals approximates a normal distribution, with slight skewness and heavy tails. Altogether, the residuals resemble white noise, supporting the appropriateness of the model for forecasting purposes.



Plot the auto model and the original data together

Now select my own models.

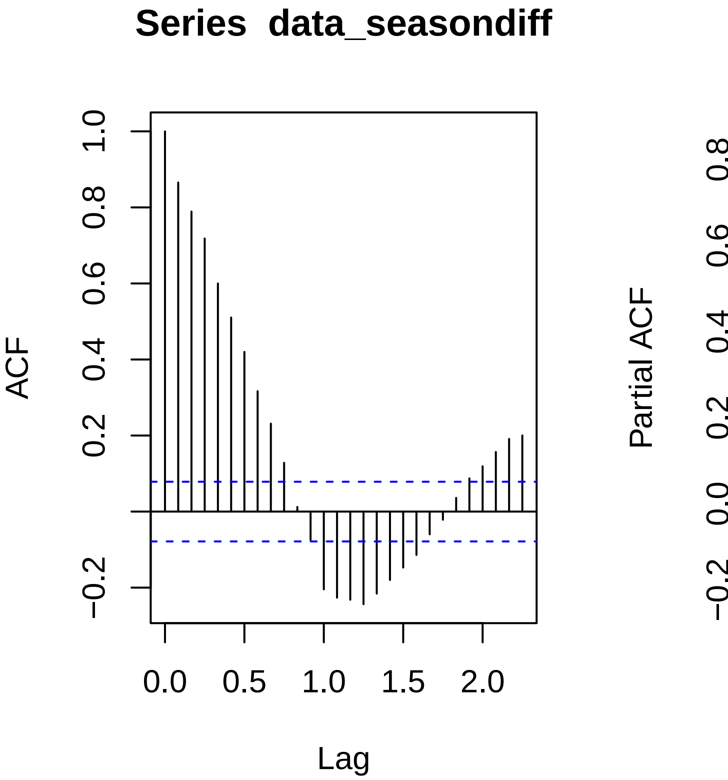
Because that I already use the test that series need seasonal differencing

```
## [1] 1
```

```
## Warning in adf.test(data_seasondiff): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_seasondiff
## Dickey-Fuller = -7.3996, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

After differencing on the seasonal pattern with D = 1, the time series are stationary. Thus, choose D = 1, and d = 0.

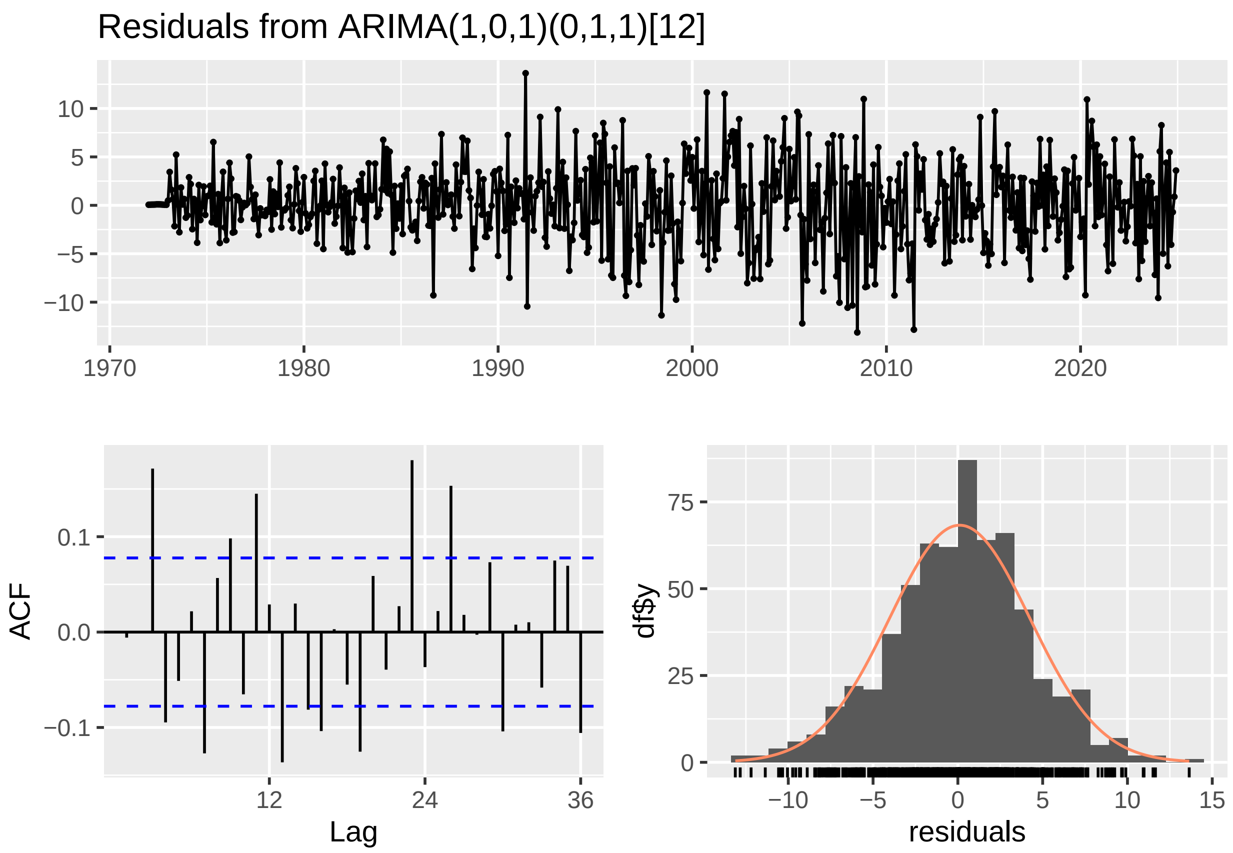


Now plot the acf and pacf of seasoned_diff time series

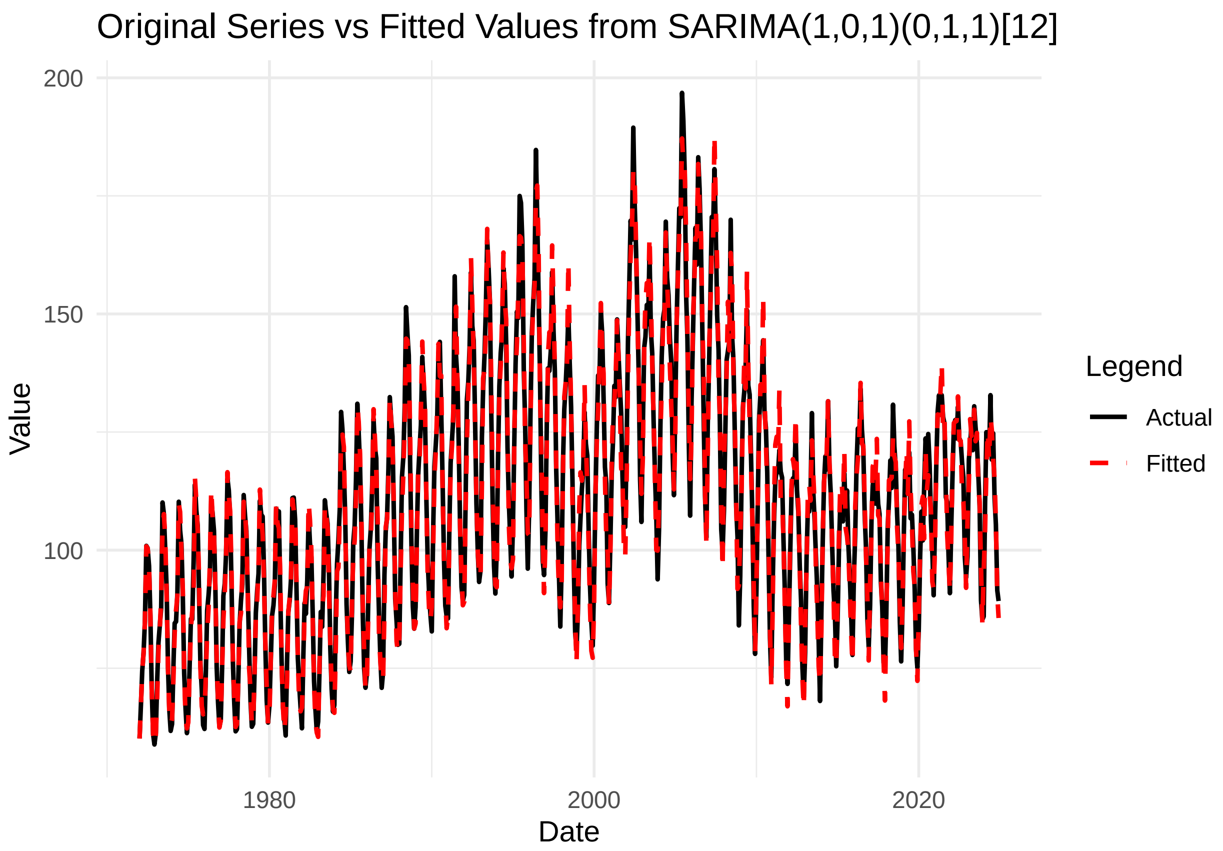
Based on the ACF and PACF plots of the seasonally differenced series (data_seasondiff), we can identify appropriate orders for a SARIMA model. The ACF shows a strong spike at lag 1 followed by a gradual decay, suggesting a non-seasonal autoregressive component, likely indicating p = 1. The PACF shows a sharp cutoff after lag 1, which supports the presence of a non-seasonal moving average component, suggesting q = 1. Since seasonal differencing has already been applied, we assume D = 1 and d = 0. While the provided plots do not extend to higher seasonal lags (e.g., 12 or 24), if further ACF analysis reveals a significant spike at lag 12, we would consider a seasonal MA term, i.e., Q = 1. If PACF shows a significant spike at seasonal lag 12 instead, a seasonal AR term P = 1 might be appropriate. In the absence of visible seasonal spikes in the current plots, a reasonable starting model would be SARIMA(1, 0, 1)(0, 1, 1)[12], with further refinement guided by residual diagnostics and information criteria such as AIC or BIC.

```
## AIC: 3573.457
```

```
## BIC: 3591.202
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1)(0,1,1)[12]
## Q* = 124.85, df = 21, p-value < 2.2e-16
##
## Model df: 3.   Total lags used: 24
```

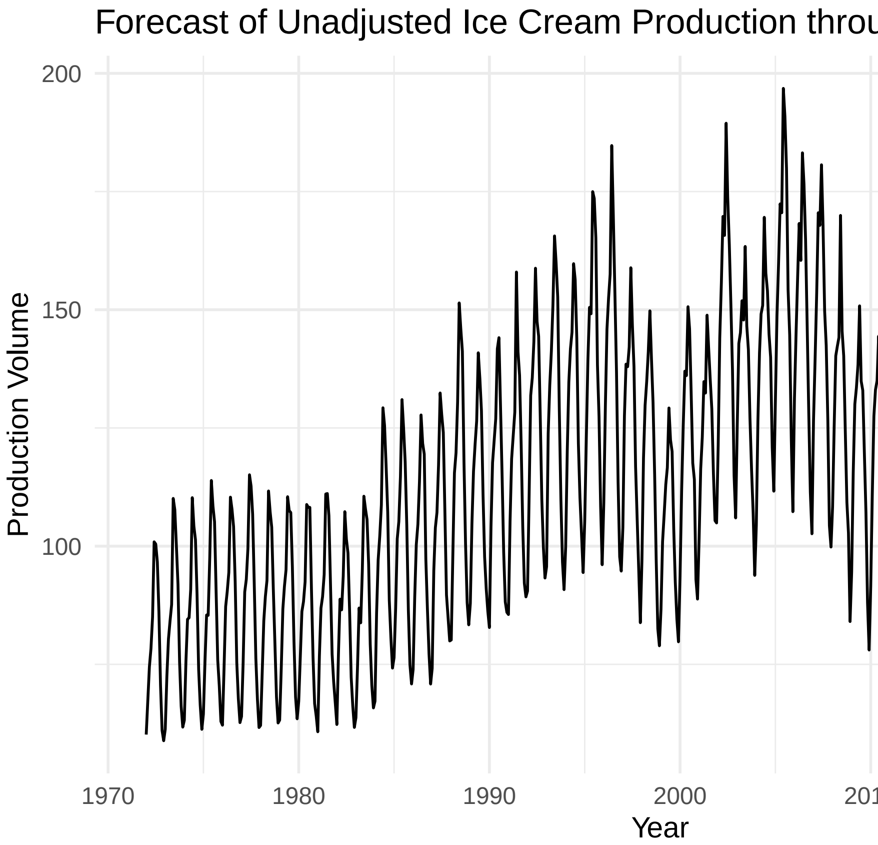


The residual diagnostics from the ARIMA(1,0,1)(0,1,1)[12] model suggest that the model provides a reasonably good fit to the time series data. The top panel shows residuals fluctuating randomly around zero without obvious patterns, indicating that the model has effectively captured the underlying structure. The ACF plot in the lower left shows that most autocorrelation values fall within the 95% confidence bounds, with no significant lag spikes, suggesting that the residuals are approximately white noise and that no substantial autocorrelation remains. The histogram in the lower right reveals a fairly symmetric, bell-shaped distribution of residuals, closely following a normal distribution, which supports the assumption of Gaussian errors. Together, these diagnostics provide evidence that the ARIMA(1,0,1)(0,1,1)[12] model is statistically appropriate and captures both the trend and seasonal dynamics of the original series effectively. My model is not as good as the auto.arima model.

QUESTION 4 Write down your chosen model. I choose the auto.arima model.

$$(1-\phi_1 L-\phi_2 L^2-\phi_3 L^3-\phi_4 L^4)(1-L^{12})Y_t=(1+\Theta_1 L^{12})\varepsilon_t$$

QUESTION 5 Forecast the unadjusted ice cream production series through December of 2025.



Comment on the validity of your forecast.

The validity of the forecast produced using the ARIMA(1,0,1)(0,1,1)[12] model appears strong based on diagnostic checks. The model residuals exhibit characteristics of white noise—centered around zero, uncorrelated, and approximately normally distributed—suggesting that the model has effectively captured the structure of the original time series, including its seasonal component. The forecast maintains the expected seasonal behavior, projecting higher production in summer months and lower in winter, consistent with historical trends in ice cream consumption. However, like all time series models, the forecast assumes that future patterns will follow past behavior. This introduces some limitations, especially in the long term, as external factors (e.g., economic shifts, supply chain disruptions, or climate anomalies) may alter production dynamics in unforeseen ways. Additionally, the prediction intervals widen over time, reflecting increasing uncertainty. Overall, the forecast is statistically sound and appropriate for short-to-medium-term planning, but should be used alongside other market insights for long-term strategic decisions.