STATS 4A03 Time Series Analysis Project

Time Series Forecasting of Monthly Ice Cream Sales (1972–2020)

Zhiyan Chen (400365265)

STATS 4A03: Time Series

Dr. J. E. Paguyo

April 6, 2025

Section 1: Introduction

The goal of this project is to perform a statistical analysis of monthly ice cream sales from January 1972 to January 2020. The dataset contains 577 monthly observations, although the unit of sales and the region where the data was collected are not specified. However, it has great performance in capturing trends over time and showing patterns in sales, allowing us to reflect long-term sales trend and seasonal fluctuations in ice cream demand. The data was downloaded from Kaggle, consisting of two columns: DATE (the date of observation) and IPN31152N (units sold).

In this project, we aim to fit a seasonal ARIMA model to forecast monthly ice cream sales after 2020. Since ice cream sales have historically peaked during the summer months, it is important for businesses to understand customer behaviour and plan their sales strategies accordingly. This strong seasonality also suggests that time series models which reflect both trend and seasonal patterns are effective for this type of analysis.

That is, the main objective of this project is to apply basic time series techniques to identify the most appropriate forecasting model.

Section 2: Modeling

First, the monthly sales data was converted to a time series object with a frequency of 12. A plot of the raw data is showed in *Figure 1*. We can see a clear seasonal pattern followed by an upward then relatively stable trend over time.

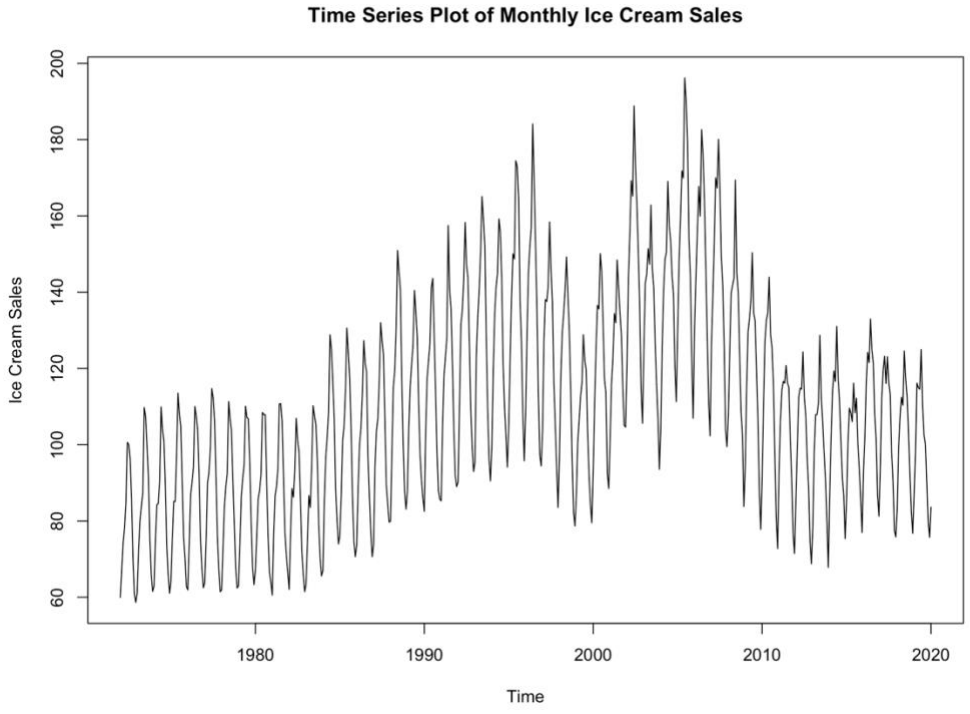

Figure 1

An Augmented Dickey-Fuller (ADF) test returned a p-value of 0.6307 (> 0.05), indicating that the original time series was not stationary. To address this, a first-order difference was applied (d =1), resulting in a series with a more stable mean. The differenced series passed the ADF test with a p-value < 0.01, confirming its stationarity. However, residual seasonality still remained, as indicated by the repeated patterns in the sample ACF and a spike at lag 12 in the sample PACF (*Figure 2*). Therefore, a seasonal difference with lag 12 (D=1) was also applied.

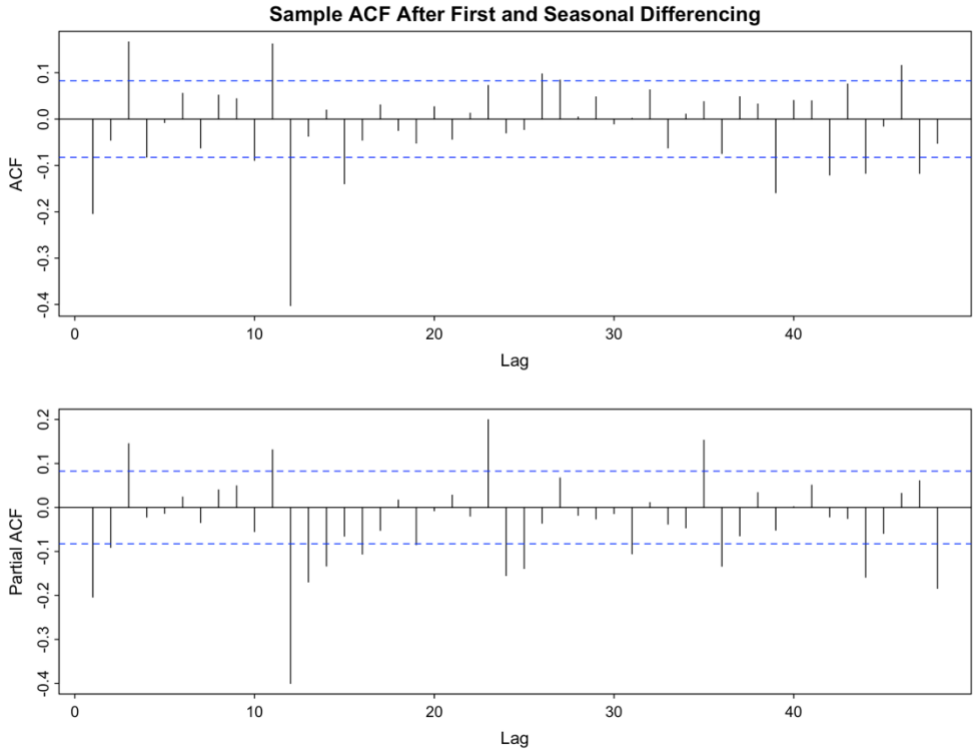

Figure 2

Figure 3 displays the sample ACF and PACF plots after applying both first and seasonal differencing to the series. In the ACF, noticeable spikes are observed at lags 1 and 12, followed by a faster decay that most other lags falling within the blue lines. This implies a short-term and seasonal autocorrelation, that is, the series has both non-seasonal and seasonal MA components (q = 1, Q = 1). The sample PACF also shows significant spikes at lags 1 and 12, as well as additional spikes at lags 24, 36, and 48, indicating a non-seasonal AR(1) and potential seasonal AR components.

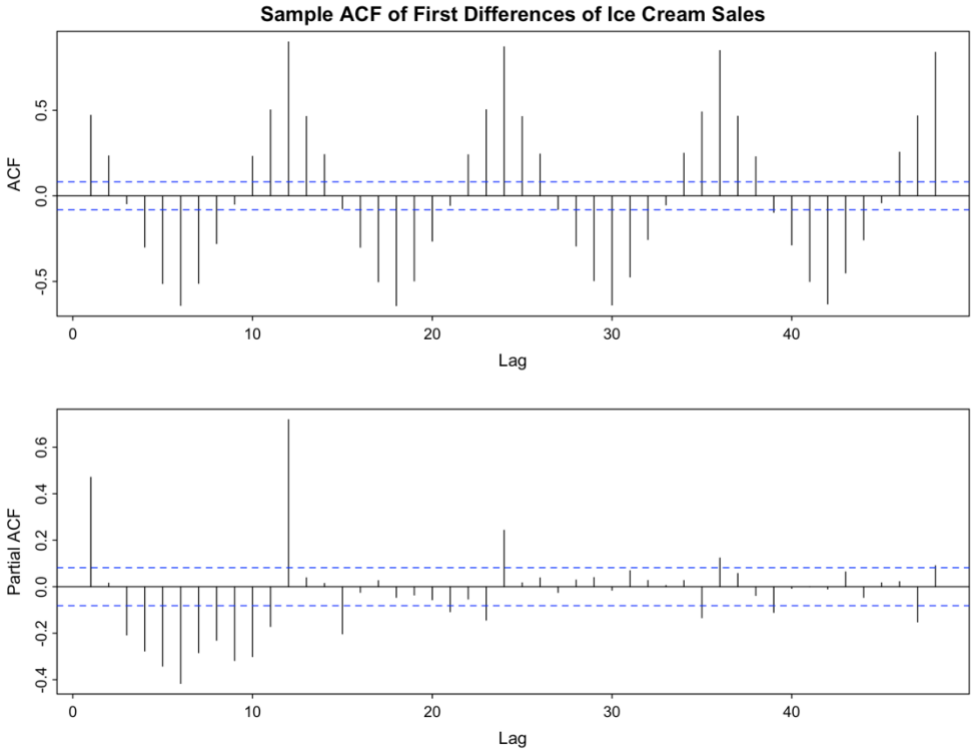

Figure 3

Based on this, we began model fitting with a basic seasonal ARIMA model: ARIMA(1,1,1)*(0,1,1)_12. This model produced a reasonably good Q-Q plot and passed the Shapiro-Wilk normality test. However, the first few Ljung-Box p-values were below 0.05, and the residual ACF indicated remaining autocorrelation. To improve the fit, we increased the non-seasonal MA order and tested ARIMA(1,1,2)*(0,1,1)_12, aiming to reduce some residual autocorrelations. The result was very similar to before. We then increased the non-seasonal AR order and fitted ARIMA(2,1,2)*(0,1,1)_12. This model produced flatter ACF and PACF plots of residuals, but the Ljung-Box p-values became worse and all remained below 0.05. We also tested different seasonal AR components (P = 0, 1, 2, 3) to confirm the spikes at multiples of 12, but they had little to no impact on the model's performance. Finally, we tested ARIMA(2,1,3)*(0,1,1)_12 by further increasing the MA order. This model produced the best overall diagnostics.

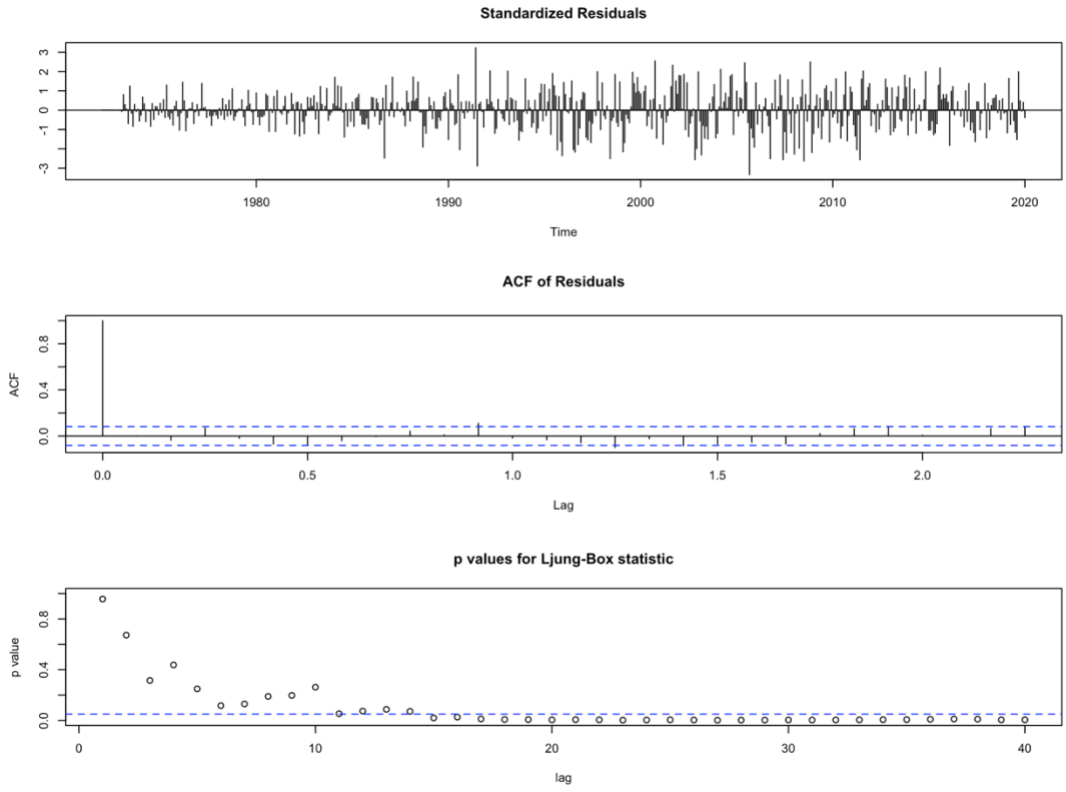

Figure 4

From *Figure 4* we can see that, the residuals appeared more like white noise, with a greater amount of Ljung-Box p-values exceeding the 0.05 threshold across multiple lags.

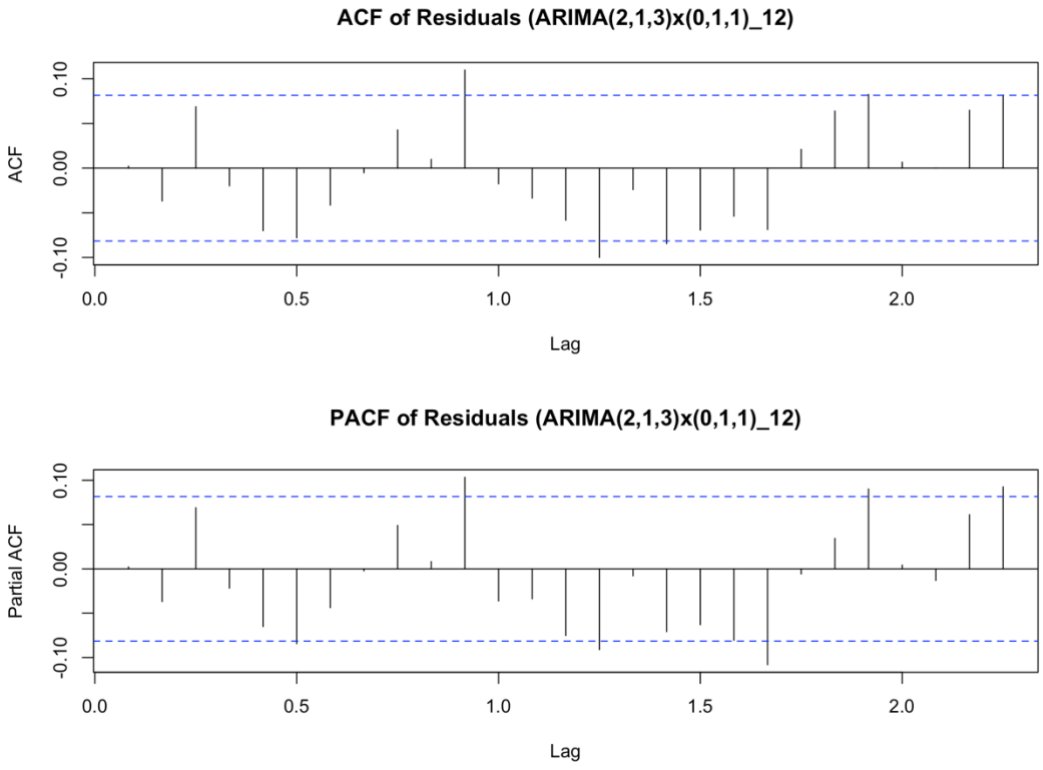

Figure 5

From *Figure 5*, the ACF and PACF plots of the residuals appear flatter than in previous models, with only a few minor spikes present. Those small spikes are acceptable as they are not statistically significant and unlikely to have a substantial impact on the model's performance.

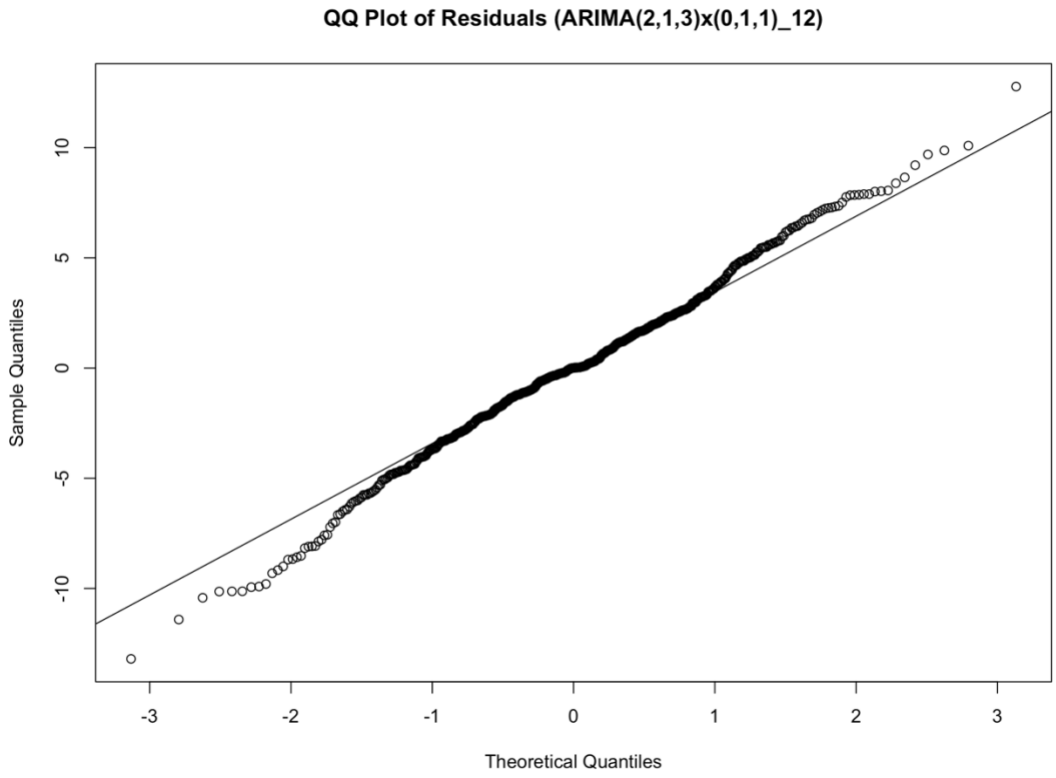

Figure 6

In *Figure 6*, we observe a few outliers in the tails of the residual distribution, similar to the earlier models. However, the Shapiro-Wilk test for normality has a test statistic of W = 0.99434, with a p-value of 0.03085, and normality is not rejected at any of the usual significance levels (Cryer & Chan, 2008, p. 240). The QQ plot in *Figure 6* also shows an improved alignment with the theoretical normal distribution compared to previous models.

As one further check on the model, we consider overfit with an ARIMA(2,1,4)*(0,1,1)_12 model. As shown in *Figure 7*, the estimates of theta_3 and Theta_1 changed very little in magnitude, with even minimal changes in size of their standard errors. Moreover, the estimate of the new parameter, theta_4, is not far from zero. The estimated sigma^2 and the log-likelihood have also not changed much while the AIC has increased (Cryer & Chan, 2008, p. 240). This suggests that the added complexity did not improve the model and our original ARIMA(2,1,3)*(0,1,1)_12 performed better. Similar process was also tested for ARIMA(2,1,2)*(0,1,1)_12 and ARIMA(2,1,3)*(0,1,1)_12 model to make sure our decision was truly the simplest and most appropriate.

```
Call:
arima(x = ts_icrm, order = c(2, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ar1      ar2     ma1     ma2      ma3     sma1
      -1.1557  -0.9966  0.9955  0.8087  -0.1404  -0.6853
s.e.   0.0040   0.0047  0.0432  0.0510   0.0429   0.0321

sigma^2 estimated as 15.64:  log likelihood = -1580.56,  aic = 3173.11
```

```
Call:
arima(x = ts_icrm, order = c(2, 1, 4), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
         ar1     ar2      ma1      ma2     ma3      ma4     sma1
      0.6642  0.1128  -0.8644  -0.0310  0.2152  -0.2017  -0.7045
s.e.  0.2059  0.2776   0.2026   0.2915  0.0664   0.0449   0.0386

sigma^2 estimated as 16.36:  log likelihood = -1592.79,  aic = 3199.57
```

Figure 7

Section 3: Result

As a result, the final model, ARIMA(2,1,3)*(0,1,1)_12, was used to generate a 24-month forecast of monthly ice cream sales. *Figure 8* displays the last few years of the monthly ice cream sales time series together with forecasts and 95% forecast limits for two additional years (Cryer & Chan, 2008, p. 205). The forecasts follow the approximate cycle in the actual series and the forecast limits are quite close to the fitted trend forecast. Comparing the 2022 forecast to 2021's, the forecast intervals widen slightly over time, which reflects the increasing uncertainty. Noted that this is normal, and the seasonal path is more valuable for us to consider this model to be useful to support local ice cream businesses.

With a clear idea of when ice cream demand is likely to increase or decrease, businesses can plan more confidently and efficiently. For instance, they can prepare a few months ahead to increase their supplies before the expected summer peaks. They can also explore other customer trends (e.g. trending flavours) during off-peak months to better prepare.

Although forecast does not account for external factors, it still supports our initial goal of understanding this long-term trend and provides some insights for businesses to consider. For example, our forecast period (2020–2022) overlaps with the global COVID-19 pandemic, during which the overall market including ice cream sales may have declined.

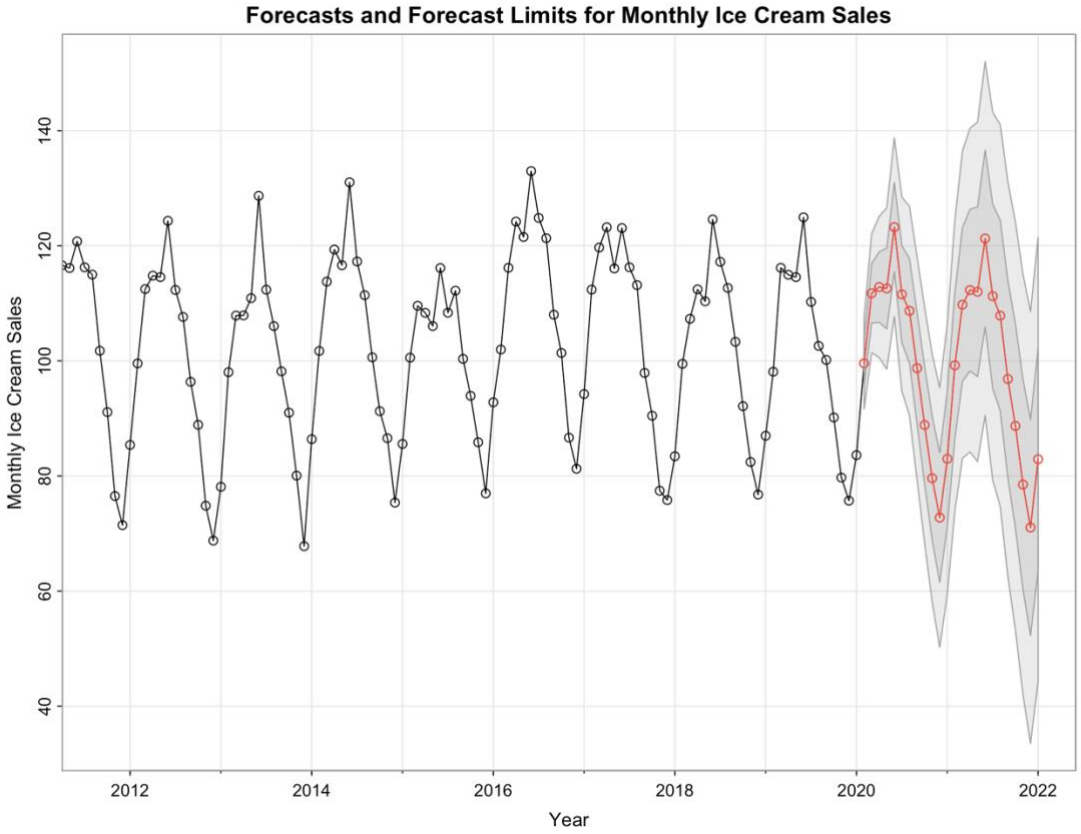

Figure 8

Section 4: Conclusion

While the final model captures the overall structure of the time series well, it has some limitations. Most importantly, as previously noted, the forecast period overlaps with a major historical event. During the pandemic, customer habits and the market conditions were significantly affected. These external factors were not included in the model, as it was only based on historical sales data.

To improve this, and addressed the missing details in original dataset, additional variables such as geographic location, weather patterns, and economic indicators could be incorporated. We might need access to more detailed raw data or collected more information from 1972 to 2020. It would also be beneficial to include post-2020 data to update the model with more recent consumer behavior and market trends.

Nevertheless, the short-term forecast successfully supports the project's original goal of analyzing long-term trends and seasonal demand in ice cream sales. This study shows the value of statistical model in guiding business planning, while also emphasizes the importance of relating to real world conditions when developing mathematical or statistical models.

Reference

Cryer, J. D., & Chan, K.-S. (2008). *Time series analysis: With applications in R* (2nd ed.).

Abd El-Ghafar, S. (n.d.). *Monthly Ice Cream Sales Data (1972–2020)*. Kaggle. https://www.kaggle.com/datasets/abdocan/monthly-ice-cream-sales-data-1972-2020