Analyzing Sales Trends and Forecasting Demand for Cooking Sprays: Identifying Influencing Factors through Statistical Modeling

Table of Contents

Table of Figures

1 INTRODUCTION

The market for cooking sprays has greatly expanded in recent years because to the growing trend toward healthy food and lifestyle choices. By concentrating on expansion and maximizing its portfolio of Cooking Spray brands, we seek to drive growth and shareholder value. In recent years, cooking sprays have grown in popularity, especially among consumers who are looking for healthier substitutes for conventional cooking techniques. These sprays are a popular option for people trying to stick to a healthy diet because they provide an easy method to flavor food while using less butter or oil when cooking.

As a result, retailers are now stocking a variety of different brands and flavors of cooking sprays to cater to this demand. By accurately predicting the demand for cooking sprays, retailers can ensure they have the right products in stock at the right time, which can help them avoid stockouts and overstocking. This can lead to increased sales and profits while reducing inventory costs.

2 SCOPE

In this project, we aim to conduct an in-depth analysis of one of the largest companies in the US to predict their future sales, which is a critical component of strategic planning. Our analysis encompasses various internal and external factors that can potentially impact on the company's weekly sales. We have collected data from 2018 to 2022, encompassing different brands across the country. Our analysis includes time series analysis for sales prediction using regression and SARIMA.

3 LITERATURE REVIEW

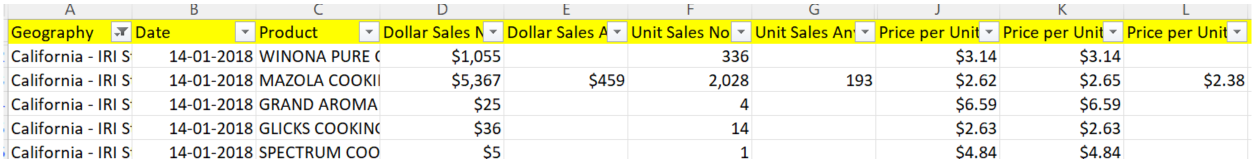
The importance of sales forecasting and inventory management has been widely recognized in the field of operations management. Previous studies have focused on various aspects of sales forecasting and inventory management, including the impact of demand variability and lead time on inventory management, as well as the use of inventory control policies such as economic order quantity (EOQ) and just-in-time (JIT) to improve inventory management performance.

However, there is a lack of research specifically focused on the cooking spray industry, despite its growing importance in the food industry. In response to this gap, our study aims to develop a sales forecasting model that can be used to effectively manage inventory and throughput in the cooking spray industry. Apart from considering external factors such as seasonality and promotions, our study aims to explore the impact of additional variables such as time and sales of the previous year on the sales of cooking sprays. Through analyzing the relationship between these factors and sales, we aim to gain deeper insights into the factors that drive demand for cooking sprays.

4 DATA

The dataset used for the analysis contains information about the total dollar sales from 2018 to 2022 for different regions across the United States. The dataset has a time series structure, with observations recorded monthly over a period of several years. The variables included in the dataset are:

- Total dollar sales: Total revenue generated by the sale of cooking sprays.
- Date: The date of each observation.
- Incremental dollar sales: Additional revenue generated by the sale of a specific product.
- Base dollar sales: Revenue generated by the sale of a product without any additional promotions or discounts.
- Price per unit: Price of a single unit of the product.
- Price per unit with merchandise: Price of a single unit of the product when sold with additional merchandise.
- Price per unit without merchandise: Price of a single unit of the product when sold without additional merchandise.
- Brands: Brand names of the product.



A	B	C	D	E	F	G	J	K	L
Geography	Date	Product	Dollar Sales N	Dollar Sales A	Unit Sales No	Unit Sales An	Price per Unit	Price per Unit	Price per Unit
California - IRI S	14-01-2018	WINONA PURE (	$1,055		336		$3.14	$3.14	
California - IRI S	14-01-2018	MAZOLA COOKI	$5,367	$459	2,028	193	$2.62	$2.65	$2.38
California - IRI S	14-01-2018	GRAND AROMA	$25		4		$6.59	$6.59	
California - IRI S	14-01-2018	GLICKS COOKIN(	$36		14		$2.63	$2.63	
California - IRI S	14-01-2018	SPECTRUM COO	$5		1		$4.84	$4.84	

Figure 1: Data Overview

5 EMPIRICAL METHOD

Empirical methods involve using statistical techniques to analyze data and understand the relationships between variables. This approach typically involves collecting data and applying statistical methods such as regression analysis and time series analysis to uncover patterns and insights. We are using empirical methods to identify the internal and external factors influencing sales.

5.1 Trend Analysis

For the trend regression of the cooking sprays dataset, we can use the ordinary least squares (OLS) method to estimate the relationship between the dependent variable (sales) and the independent variables (price, time). The estimating equation for the trend regression can be written as:

$$\text{Sales} = \beta 0 + \beta 1\text{Price} + \beta 2\text{Time} + \varepsilon$$

Where,
Sales is the dependent variable representing total unit sales.
Price is the independent variable representing the price per unit.
Time is the independent variable representing the time in months from the start of the dataset.
$\beta 0$, $\beta 1$, and $\beta 2$ are the coefficients to be estimated by the OLS method
ε is the error term representing the unobserved factors that affect the dependent variable but are not included in the model.

In addition to the estimating equation, we can also perform statistical tests to determine the significance of the independent variables in explaining the variation in the dependent variable. The t-test can be used to test the null hypothesis that the coefficient of an independent variable is equal to zero. The F-test can be used to test the overall significance of the model, i.e., whether at least one of the independent variables is significantly related to the dependent variable.

5.2 AR Regression

For the time series analysis of the cooking sprays dataset, we first converted the date variable into a time series object and then plotted the time series to visually explore any trends, seasonality, or other patterns in the data. We also perform a decomposition of the time series to separate out the trend, seasonal, and random components of the data.

For the AR regression analysis, we use the autoregressive model to capture the effect of the previous sales on the current sales. The estimating equation for the AR(p) model is as follows:

$$y(t) = c + \varphi 1y(t-1) + \varphi 2y(t-2) + ... + \varphi p*y(t-p) + \varepsilon(t)$$

where $y(t)$ represents the sales at time t, c is the intercept term, $\varphi 1, \varphi 2, ..., \varphi p$ are the autoregressive coefficients representing the effect of the previous sales on the current sales up to p lags, and $\varepsilon(t)$ is the error term.

To estimate the parameters of the AR model, we use the maximum likelihood estimation (MLE) method. The MLE method seeks to find the values of the autoregressive coefficients that maximize the likelihood of observing the given sales data. We also check for the stationarity of the time series data and ensure that the residuals are white noise using diagnostic tests such as the Ljung-Box test and the ACF plot.

5.3 SARIMA Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a commonly used time series model that extends the ARIMA model to include seasonality. The SARIMA model considers the seasonal component of the time series by including additional seasonal parameters in addition to the ARIMA parameters.

The estimating equation for the SARIMA model can be written as:

y_t = c + φ_1 y_t-1 + … + φ_p y_t-p + θ_1 ε_t-1 + … + θ_q ε_t-q + φ_s (y_t-s - y_t-s-1) + … + φ_ps (y_t-ps - y_t-ps-1) + ε_t

where:
y_t is the time series variable at time t
c is a constant term
φ_1, …, φ_p are the autoregressive parameters
θ_1, …, θ_q are the moving average parameters
ε_t is the error term at time t
s is the seasonal period (e.g., s=12 for monthly data with annual seasonality)

φ_s, …, φ_ps are the seasonal autoregressive parameters

The SARIMA model can be estimated using maximum likelihood estimation (MLE) to determine the values of the parameters that best fit the data. The AIC and BIC criteria can be used to select the optimal SARIMA model based on the goodness of fit and complexity of the model.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$ ---------- AR Model (1)

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$ ---------- MA Model (2)

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$ ---------SARIMA Model (3)

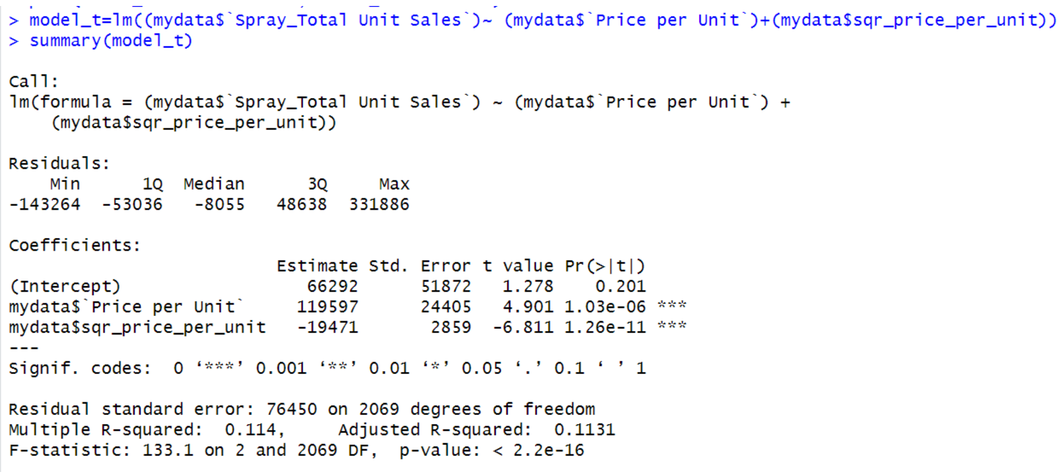
6 RESULTS

Below are the outcomes of the empirical methods:

6.1 Trend Analysis

Based on the regression equation provided below, we can make the following observations: The relationship between price and sales is concave, which implies that sales increase with an increase in price, but the trend becomes negative after a certain point.

$$\text{Sales} = 66292 + 119597\text{Price} - 19471\text{Price}^2$$



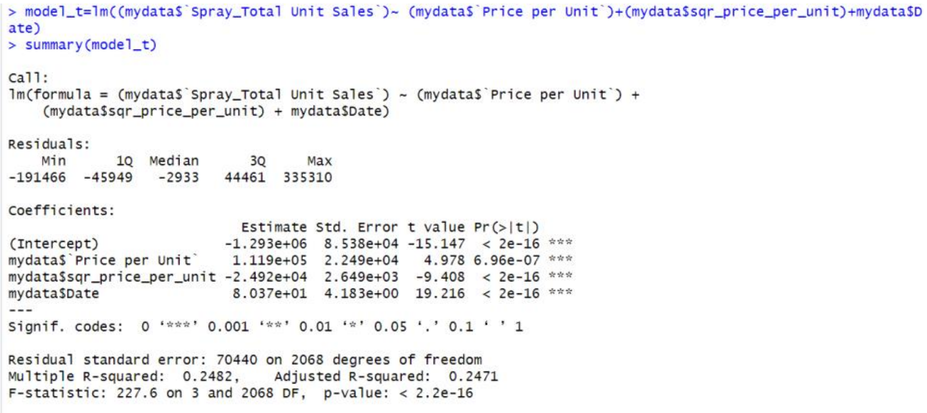
```
> model_t=lm((mydata$`Spray_Total Unit Sales`)~ (mydata$`Price per Unit`)+(mydata$sqr_price_per_unit))
> summary(model_t)

Call:
lm(formula = (mydata$`Spray_Total Unit Sales`) ~ (mydata$`Price per Unit`) +
    (mydata$sqr_price_per_unit))

Residuals:
    Min      1Q  Median      3Q     Max
-143264  -53036   -8055   48638  331886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  66292      51872   1.278    0.201
mydata$`Price per Unit`     119597      24405   4.901 1.03e-06 ***
mydata$sqr_price_per_unit   -19471       2859  -6.811 1.26e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76450 on 2069 degrees of freedom
Multiple R-squared:  0.114,	Adjusted R-squared:  0.1131
F-statistic: 133.1 on 2 and 2069 DF,  p-value: < 2.2e-16
```

Figure 2: Regression Analysis for Sales vs Price

On adding the time variable into the regression, the regression equation is as follows:

$$\text{Sales} = 1.119*10^5\text{Price} - 2.49*10^4\text{Price}^2 + 80.37\text{Time} - 1.29*10^6$$

This indicates that there is a spurious regression problem indicating a relationship between two or more unrelated time series processes simply because each has a trend and to account for the trending behavior, we added a time trend. The p-value indicates that all the above independent variables are statistically significant.

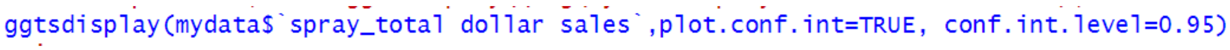


```
> model_t=lm((mydata$`Spray_Total Unit Sales`)~ (mydata$`Price per Unit`)+(mydata$sqr_price_per_unit)+mydata$D
ate)
> summary(model_t)

Call:
lm(formula = (mydata$`Spray_Total Unit Sales`) ~ (mydata$`Price per Unit`) +
    (mydata$sqr_price_per_unit) + mydata$Date)

Residuals:
    Min      1Q  Median      3Q     Max
-191466  -45949   -2933   44461  335310

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.293e+06  8.538e+04 -15.147  < 2e-16 ***
mydata$`Price per Unit`    1.119e+05  2.249e+04   4.978 6.96e-07 ***
mydata$sqr_price_per_unit -2.492e+04  2.649e+03  -9.408  < 2e-16 ***
mydata$Date                8.037e+01  4.183e+00  19.216  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70440 on 2068 degrees of freedom
Multiple R-squared:  0.2482,	Adjusted R-squared:  0.2471
F-statistic: 227.6 on 3 and 2068 DF,  p-value: < 2.2e-16
```

Figure 3: Regression Analysis including Time Trend

6.2 AR Regression

Prior to fitting the AR model, we conducted a preliminary analysis to determine the significant number of lags to include in the model. Specifically, we generated a partial autocorrelation function (PACF) plot using the below syntax in R-Studio and observed that the last lag that exceeds the confidence interval was 2.



```
ggtsdisplay(mydata$`spray_total dollar sales`,plot.conf.int=TRUE, conf.int.level=0.95)
```

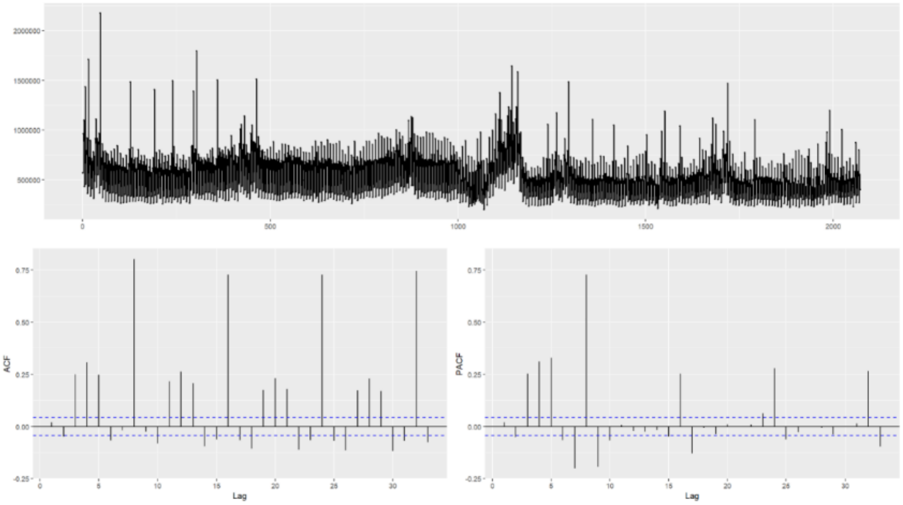
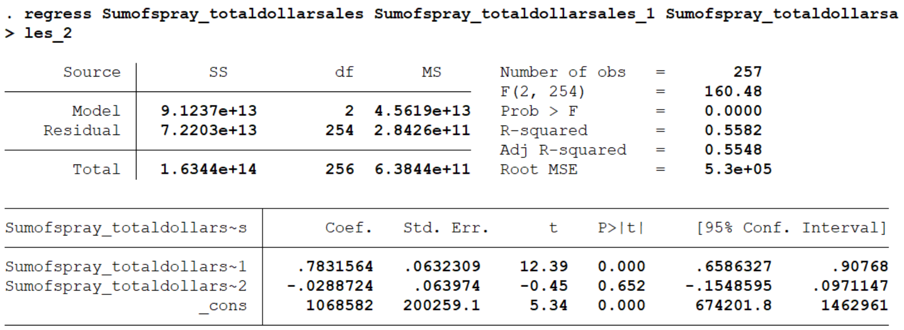


Figure 4: Plot of ACF and PACF

Revenue(t) = 1068582 + 0.78Revenue(t-1) – 0.028Revenue(t-2)



```
. regress Sumofspray_totaldollarsales Sumofspray_totaldollarsales_1 Sumofspray_totaldollarsa
> les_2

      Source |       SS           df       MS      Number of obs   =       257
-------------+----------------------------------   F(2, 254)       =    160.48
       Model |  9.1237e+13         2  4.5619e+13   Prob > F        =    0.0000
    Residual |  7.2203e+13       254  2.8426e+11   R-squared       =    0.5582
-------------+----------------------------------   Adj R-squared   =    0.5548
       Total |  1.6344e+14       256  6.3844e+11   Root MSE        =    5.3e+05

------------------------------------------------------------------------------
Sumofspray_totaldollars~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Sumofspray_totaldollars~1 |   .7831564   .0632309    12.39   0.000     .6586327      .90768
Sumofspray_totaldollars~2 |  -.0288724    .063974    -0.45   0.652    -.1548595    .0971147
                    _cons |    1068582   200259.1     5.34   0.000     674201.8     1462961
------------------------------------------------------------------------------
```

Figure 5: Trend Regression

The R-squared value of the regression is 55%, indicating that 55% of the variation in the independent variable can be explained by the given dependent variables in the model. The p-value for the second lag is greater than 0.05, which means it is not statistically significant in the above regression. The coefficient of 0.78 for the first lag (Revenue(t-1)) suggests that a one-unit increase in the value of Revenue(t-1) will result in a 0.78-unit increase in the value of Revenue(t), while all other variables remain constant.

6.3 SARIMA Model

6.3.1 Exploring Time-Series through Visualization

We created a time series plot to examine any trends, seasonal patterns, or other features that may need to be considered in the model. However, no stationary trend was observed in the plot. Therefore, we proceeded to implement the following steps to transform the time series into a stationary one.

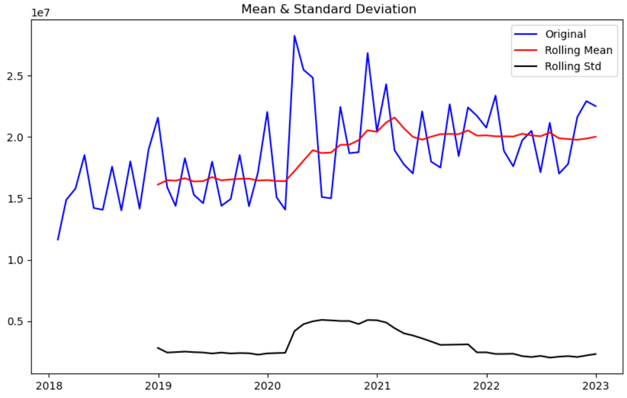


Figure 6: Time series plot for Revenue

6.3.2 Addressing Stationarity Issues in Time-Series Data

Similar to ARIMA, SARIMA also assumes that the time series is stationary. In cases where the time series is not stationary, it can be made stationary by applying techniques like differencing or logarithmic transformation. After applying logarithmic transformation to the time series, we observe that the resulting series is similar in terms of stationarity to the original one.

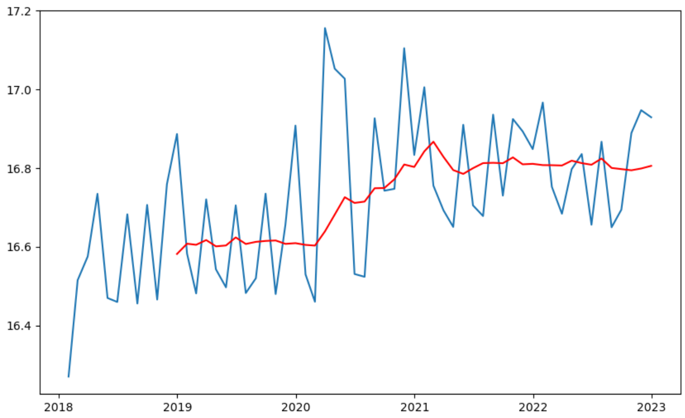


Figure 7: Time Series Plot Upon Log Transformation

6.3.3 Determine the order of differencing

To make a time series stationary, it's necessary to determine the order of differencing for both the seasonal and non-seasonal components. Once the differencing is applied, the data is decomposed and transformed to achieve stationarity. This has removed the time trend, seasonality in the data.

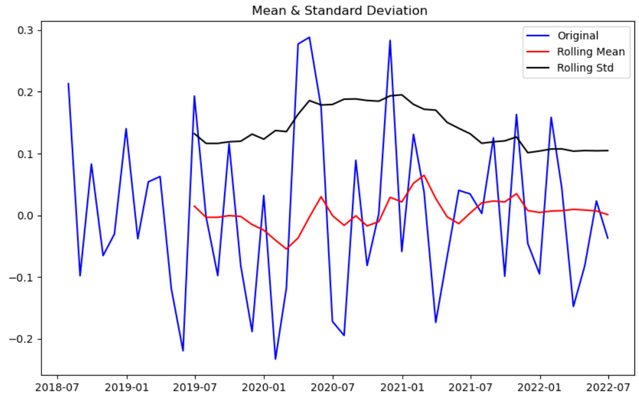


Figure 8D: Time Series Plot Upon Differencing

6.3.4 Determining Trend and Seasonal Components of Time-Series

The trend and seasonal components have been identified below after the application of the order of differencing.

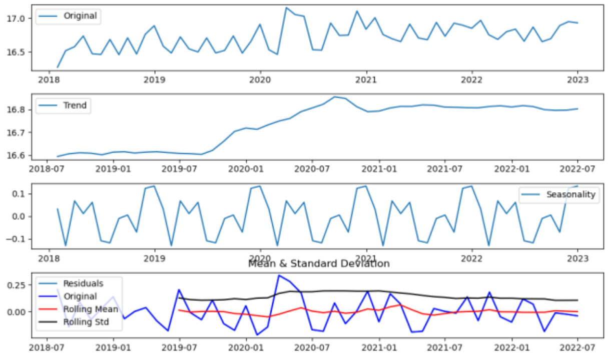
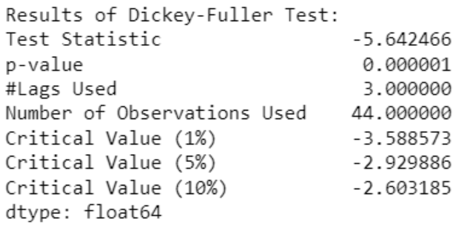


Figure 9: Seasonality, Trend Plots

6.3.5 Evaluating Stationarity using the Dicky-Fuller Test

We employed the Dickey-Fuller test to confirm the stationarity of the data, with the null hypothesis (H0) being that the data is stationary. The results reveal that the data is indeed stationary, as presented below.

The p-value of less than 0.05 from the test summary indicates that the data is stationary.



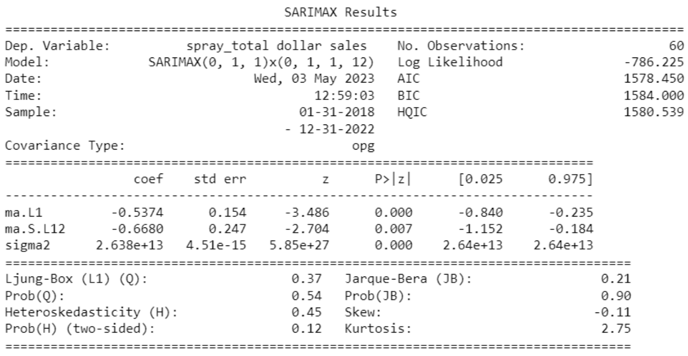
```
Results of Dickey-Fuller Test:
Test Statistic                 -5.642466
p-value                         0.000001
#Lags Used                      3.000000
Number of Observations Used    44.000000
Critical Value (1%)            -3.588573
Critical Value (5%)            -2.929886
Critical Value (10%)           -2.603185
dtype: float64
```

Figure 10: Dickey-Fuller Test Result Summary

6.3.6 Building a SARIMA Model for Time-Series Forecasting

Fitted the SARIMA model to the time series data using the chosen AR, MA, and seasonal parameters and the results is shown below.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$



```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:          spray_total dollar sales    No. Observations:                   60
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 12)   Log Likelihood                -786.225
Date:                            Wed, 03 May 2023   AIC                           1578.450
Time:                                    12:59:03   BIC                           1584.000
Sample:                                01-31-2018   HQIC                          1580.539
                                     - 12-31-2022                                         
Covariance Type:                              opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.5374      0.154     -3.486      0.000      -0.840      -0.235
ma.S.L12      -0.6680      0.247     -2.704      0.007      -1.152      -0.184
sigma2      2.638e+13   4.51e-15   5.85e+27      0.000    2.64e+13    2.64e+13
===================================================================================
Ljung-Box (L1) (Q):                   0.37   Jarque-Bera (JB):                 0.21
Prob(Q):                              0.54   Prob(JB):                         0.90
Heteroskedasticity (H):               0.45   Skew:                            -0.11
Prob(H) (two-sided):                  0.12   Kurtosis:                         2.75
===================================================================================
```

From the above SARIMAX results, it can be inferred that

$$\mathbf{Revenue_t = \mu + \varepsilon t - 0.5374\varepsilon_{t-1} - 0.6680\varepsilon_{t-2}}$$

Where, the non-seasonal MA coefficient is ma.L1 = -0.5374 and the seasonal MA coefficient is ma.S.L2 = -0.6680.

Upon analyzing the SARIMA plot, it can be observed that there is an upward trend in the revenue of cooking sprays.

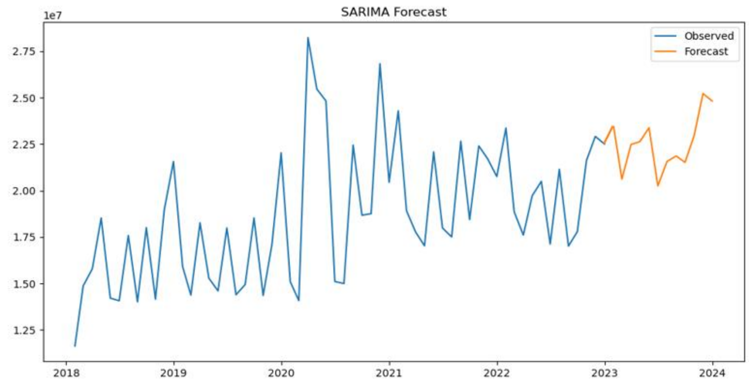


Figure 11: Forecasting Plot

7 CONCLUSION

Based on the trend analysis, we observed an increasing trend in the revenue of cooking sprays. The AR model results indicate that price has a positive impact on sales, and the trend becomes negative after a certain turning point. Meanwhile, the SARIMA model suggests that time trend and seasonality are significant factors affecting sales. We identified the order of differencing required to make the time series stationary, both for the non-seasonal and seasonal components. The model indicates that there is an increasing trend in cooking sprays revenue. Therefore, we can conclude that the factors affecting the sales include price, time trend, and seasonality, and the trend analysis and forecasting suggest an increasing trend in cooking sprays revenue.