PHASE 4 PROJECT PRESENTATION

Presented by Khadija Omar,Trevor Mwangi and Stephen Kimiti.

Contents

Project Overview

The goal of the project is to forecast upcoming monthly sales data for a Shampoo Company named Hair Haven, to help plan out the supply with demand for the company.

Firstly, we investigated and prepared the time series data. The provided data was appropriate to use time series models and we held out the last 4 periods of data points for validation.

Project Overview

Then, we determined Trend, Seasonal, and Error components in the data based on decomposition plots. After that, we analyzed the data by applying various modeling techniques and described the errors for the models. We compared the in-sample error measurements to the models, compared their AIC and BIC values, and compared error measurements for the holdout sample in the forecast. Finally, we chose the best-fitting model and forecast the next four periods.

Business Understanding

Hair Haven operates in the competitive market of hair care products, specializing in shampoos. To thrive in this dynamic industry, Hair Haven needs to implement a robust forecasting system to anticipate and adapt to changes in customer demand, market trends, and external factors.

By implementing a comprehensive forecasting strategy, Hair Haven can navigate the complexities of the hair care market, optimize operations, and stay ahead in an ever-changing industry. The insights gained from accurate monthly sales forecasts will empower the company to make data driven decisions, enhance customer satisfaction, and achieve sustainable growth.

Business Understanding

The primary goal of this project is to provide a forecast for the next 4 months of sales for the Hair Haven brand and report the findings.

Specific Objectives:
1. To forecast sales using time series modeling.
2. Do exploratory data analysis on the data.
3. Utilize time series analysis techniques to identify underlying patterns, trends, and seasonality in the monthly sales data.
4. Build a time series predictive model that can forecast monthly sales.
5. Evaluate the forecasting model performance by comparing its predictions against actual sales.

Data Understanding

The provided dataset represents monthly sales data for the Hair Haven shampoo brand. Each row corresponds to a specific month, and the associated monthly sales figures are given. Let us break down the key aspects of the data:

Time: The data spans from January 2008 to September 2013, covering multiple years. This extended period allows for the identification of seasonal patterns and trends.

Monthly Sales: Monthly sales figures are provided in the "Monthly Sales" column. Sales figures vary from month to month, reflecting the dynamic nature of the hair care product market.

Data Preparation

In this section, we are going to do several actions to prepare our data for exploratory data analysis and modeling. We will import all the necessary libraries, load the dataset using the pandas library, preview the data, do exploratory data analysis, conduct data preprocessing, and determine Trend, Seasonal, and Error components by decomposition.

EDA Analysis

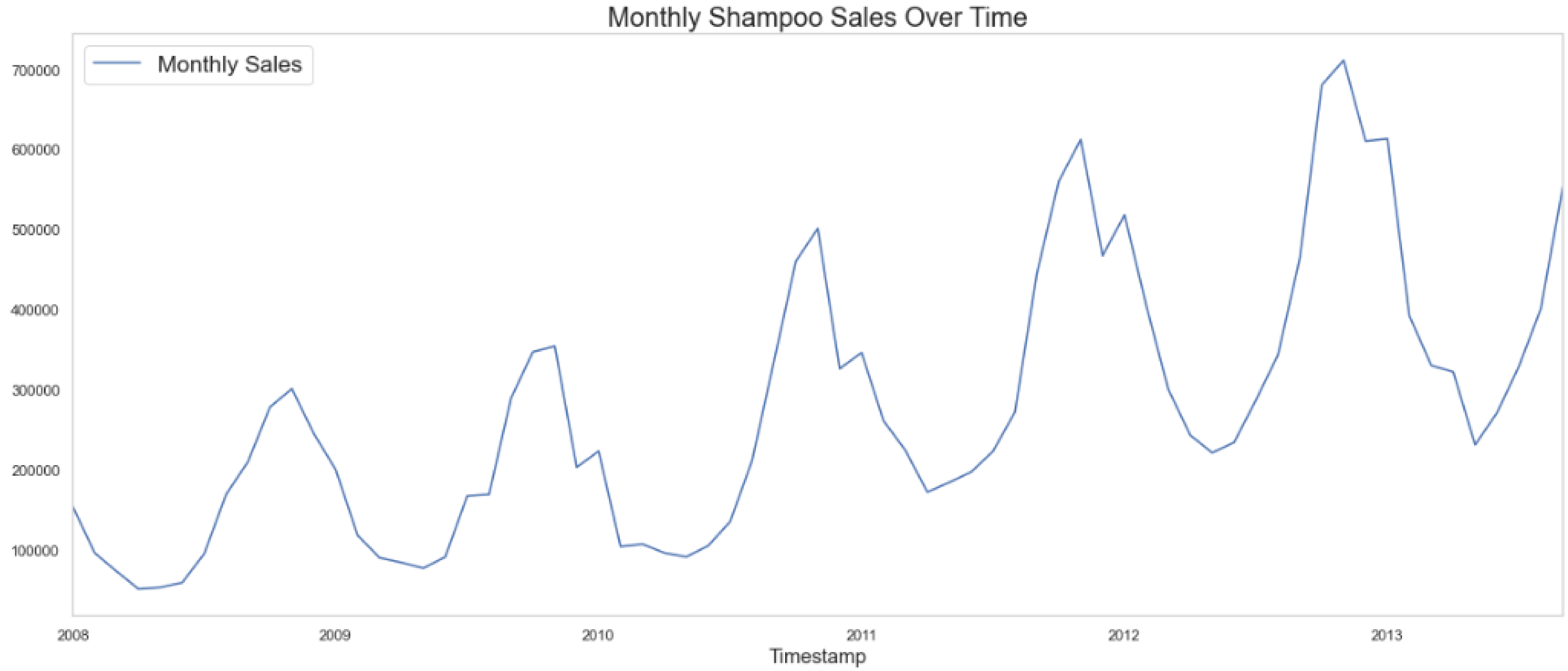
Following EDA Analysis:
Time Series Plots of Sales Amount shows the general movement of sales data. The monthly sales amount of the company is generally increasing over time. There seems to be a trend (increasing).

The yearly box plot suggests that the series has a significant trend. Sales are increasing then reducing in the year 2013, the presence of outliers in this year may represent months with exceptionally high sales compared to the typical pattern.

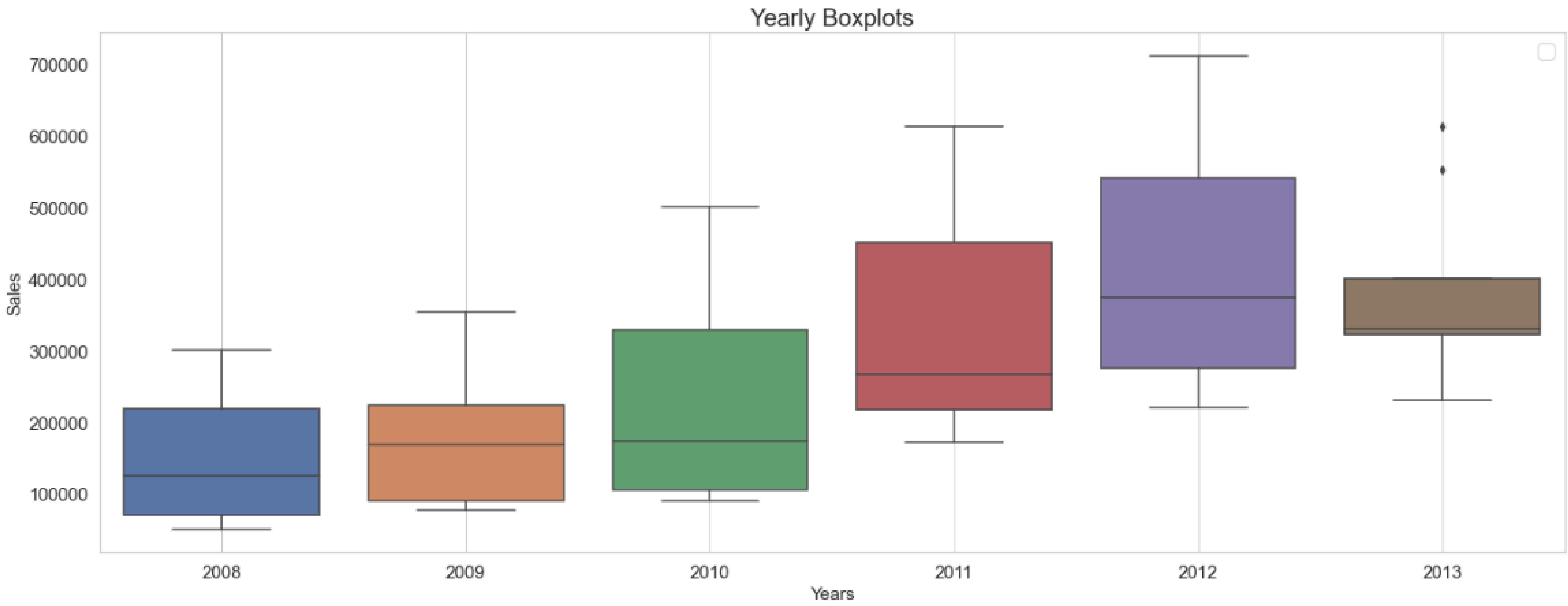
Monthly box plots suggest that we have a seasonal component each year. Also, we see that there are no outliers present.

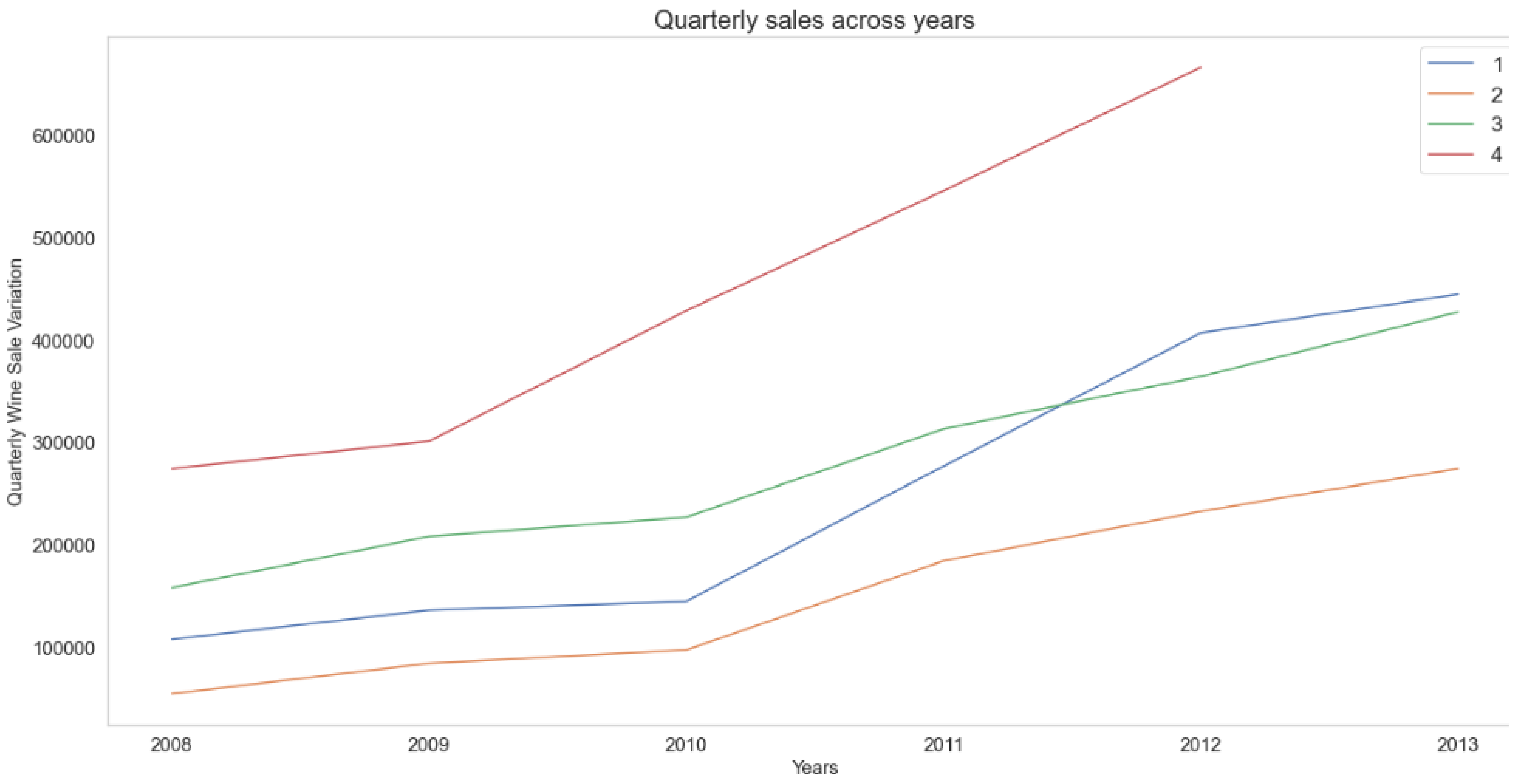
As seen from plotting quarterly sales across years, Q4 has the highest sales. After that Q3, Q1 then Q2 have the lowest sales.

Plotting sales for every year shows that the average monthly sales generally increase every year then there is a slight decrease in the year 2013.



Monthly Shampoo Sales Over Time
Monthly Sales
700000
600000
500000
400000
300000
200000
100000
2008
2009
2010
2011
2012
2013
Timestamp


Yearly Boxplots
700000
600000
500000
400000
300000
200000
100000
Sales
2008
2009
2010
2011
2012
2013
Years



Quarterly sales across years
Quarterly Wine Sale Variation
Years
100000
200000
300000
400000
500000
600000
2008
2009
2010
2011
2012
2013
1
2
3
4

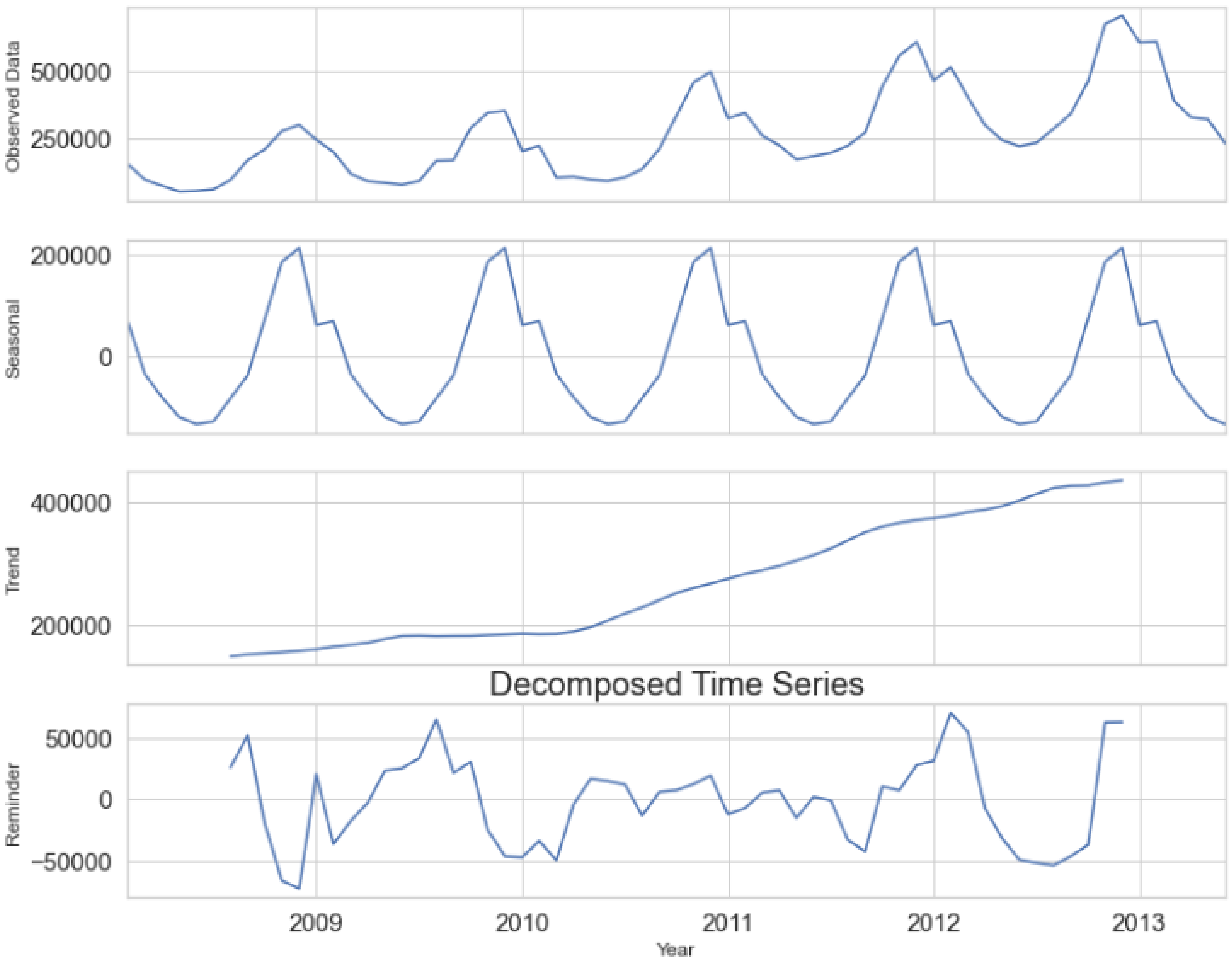
Data Preparation

Train & Test records: In preparation for the construction of predictive models, we filter out the last 4 records (from 2013-06-01 to 2013-09-30), as a holdout sample so that we can check the accuracy of my model to forecast predicted values against the actual values.

Determining Trend, Seasonal, and Error Components: The decomposition plot shows our time series are broken down into three components: trend, seasonal, and error. Each of these components makes up our time series and helps us confirm what we saw in our initial time series plot.

The Trend line confirmed that there is an upward trend.

The Seasonality subplot shows that the regularly occurring spike in sales each year changes in magnitude, ever so slightly. Our dataset contains seasonality.

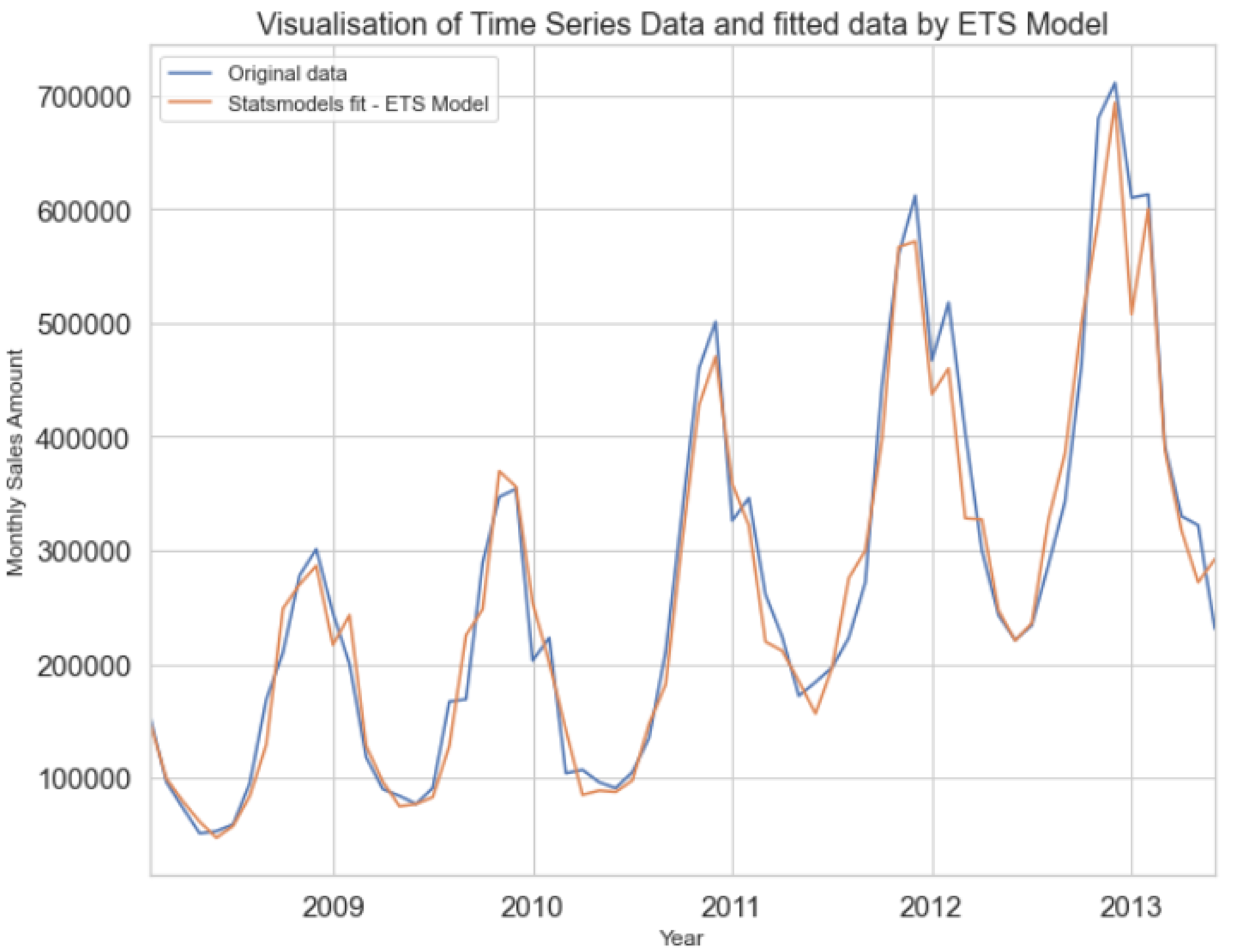


Observed Data
500000
250000
Seasonal
200000
0
Trend
400000
200000
Decomposed Time Series
Reminder
50000
0
-50000
2009
2010
2011
2012
2013
Year

Modelling

We analyze the decomposition graphs to inform forecasting models on the business problem. In this section, we determine the appropriate measurements to apply to the ETS model, the (Seasonal) ARIMA, ARMA, and ARIMA models

We start by fitting the ETS model. The error plot of the decomposed graph presents fluctuations between large and smaller errors as the time series goes on. Since the fluctuations are not consistent in magnitude then we will apply error in a multiplicative manner for our ETS model.

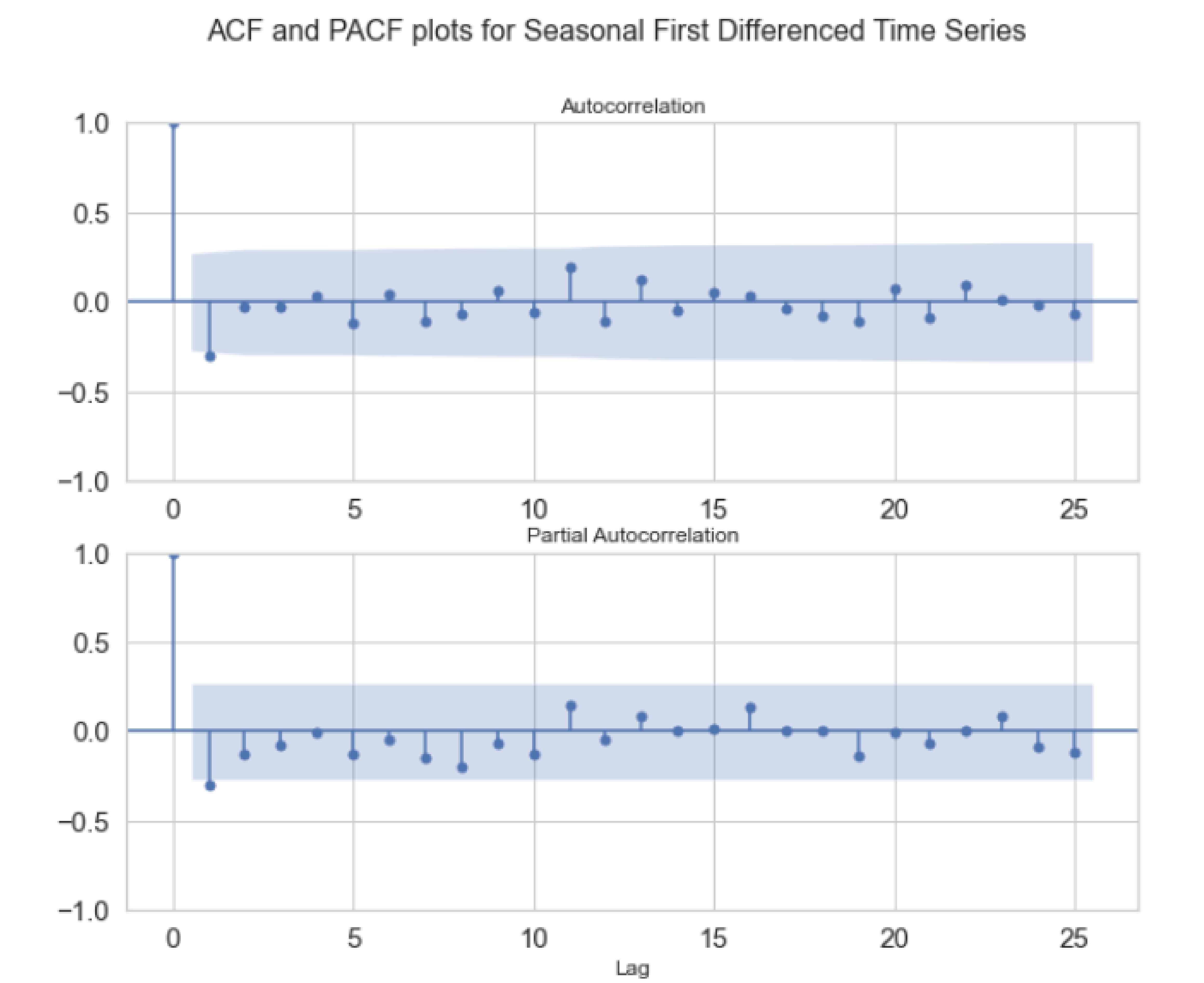


Visualisation of Time Series Data and fitted data by ETS Model
Original data
Statsmodels fit - ETS Model
700000
600000
500000
400000
300000
200000
100000
Monthly Sales Amount
2009
2010
2011
2012
2013
Year

Modelling

As the decomposition plots exhibit, the provided dataset seems to have seasonality and this suggests that any ARIMA models used for analysis will need seasonal differencing before using it for modeling. The augmented Dickey-Fuller test reminds us of the fact that the given data is not stationary.

We find the optimal parameters based on the Time Series ACF and PACF graphs. The seasonal first difference of the series has removed most of the significant lags from the ACF and PACF. It seems no need for further differencing. The remaining correlation can be considered by using autoregressive and moving average terms.


ACF and PACF plots for Seasonal First Differenced Time Series
Autocorrelation
Partial Autocorrelation
Lag
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25

Modelling

The ACF plot shows a strong negative correlation at lag-1 which is confirmed in the PACF. This suggests an MA(1) model since there is only 1 significant lag. The seasonal lags (lag 12, 24, etc.) in the ACF and PACF do not have any significant correlation so there will be no need for seasonal autoregressive or moving average terms. Therefore, the model terms for my Seasonal ARIMA model are: SARIMA(0, 1, 1)(0, 1, 0)[12]. We then fit ARMA and ARIMA models .

Evaluation

In this section, we describe the in-sample errors based on ETS models. The in-sample error measures give us a look at how well our model can predict future values. Among the various Error Terms chosen:
RMSE (Rooted Mean Squared Error)
MAE (Mean Absolute Error)
MAPE (Mean Absolute Percentage Error)
MASE (Mean Absolute Scaled Error)
Two key measures we have to check are the RMSE, which shows the in-sample standard deviation, and the MASE which can be used to compare forecasts of different models. A lower RMSE is preferred, MSE value falling well below the generic 1.00, and the commonly accepted MASE threshold for model accuracy is preferred.

We also check the AIC and BIC values. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to compare the relative goodness of fit of statistical models. Lower AIC and BIC values indicate a better fit.

Evaluation

After we analyze the training part of time series data using ETS Models, Seasonal ARIMA, ARMA, and ARIMA. Then we compare the performance of each model with In-Sample Error Measures such as RMSE, MAPE, and MASE.

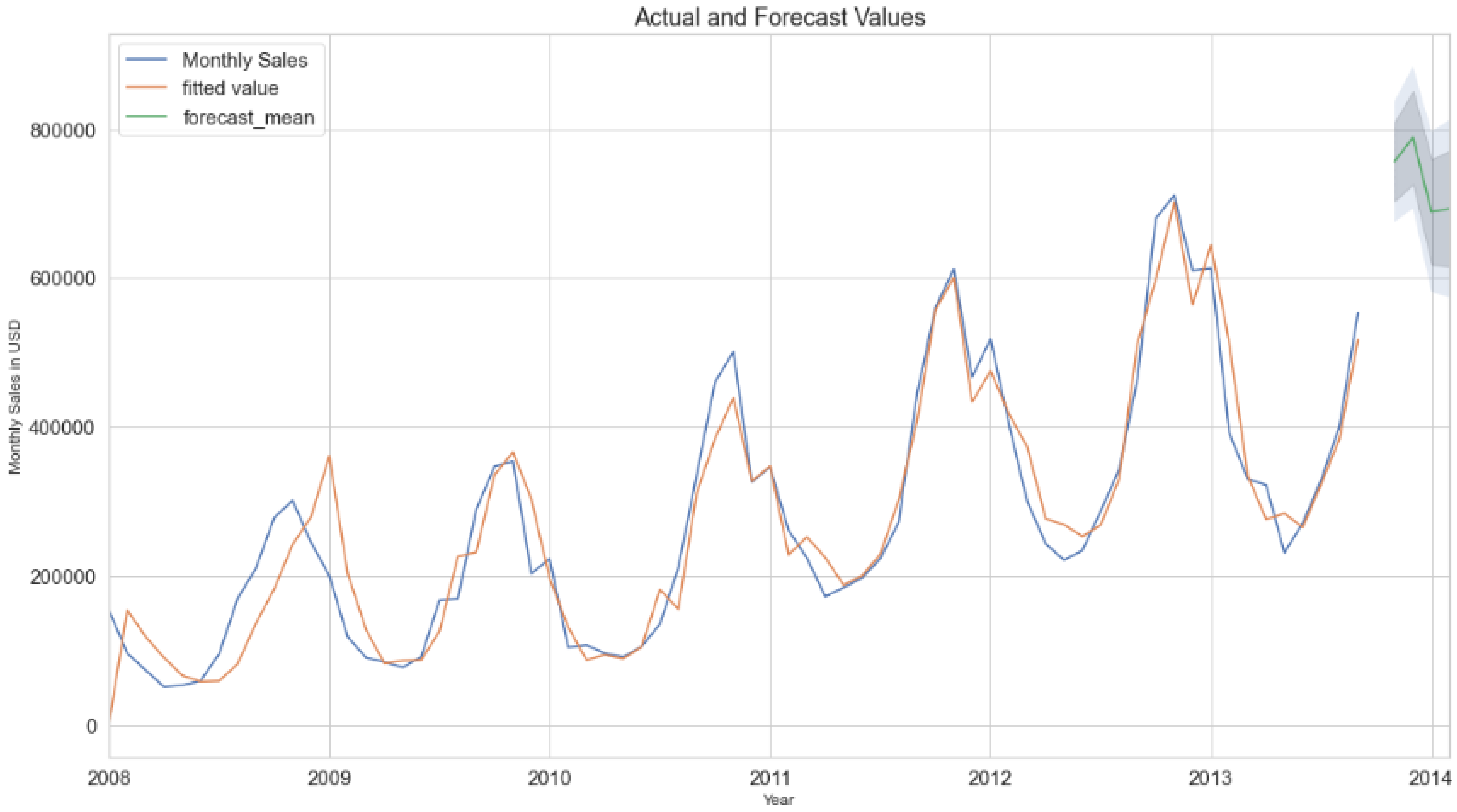
When comparing the in-sample error measures we used, the ETS model does have a much narrower standard deviation(RMSE) compared to other models. Though the MASE value of the ETS model is lower than that of Seasonal ARIMA and ARMA they are below 1.00, the generally accepted MASE threshold for model accuracy.

Based on the AIC and BIC values, the SARIMA model appears to be the most suitable model among the ones considered. It has the lowest AIC and BIC, suggesting a better trade-off between goodness of fit and model complexity.

Evaluation

Next, we compare the prediction performance of both models using holdout samples. Predicting the Holdout Sample: When looking at the model's ability to predict the holdout sample, we can recognize that the Seasonal ARIMA model shows better predictive performance in all metrics.

Forecast for the next 4 months of Sales:
Previously, we concluded that the Seasonal ARIMA model shows better performance in terms of prediction. Now, we forecast for the next four months ' sales using all the time series data based on the same Seasonal ARIMA model. Then the forecast results are calculated using 95% and 80% confidence intervals. Lastly, we visualize the forecast results



Actual and Forecast Values
Monthly Sales
fitted value
forecast_mean
800000
600000
400000
200000
0
Monthly Sales in USD
2008
2009
2010
2011
2012
2013
2014
Year

Recommendations

1. Implement the seasonal arima model.
2. Analyze the identified trends, seasonal patterns, and other insights from the model to inform marketing strategies.
3. When making business decisions based on the forecasted sales, consider the 95% confidence intervals provided by the Seasonal ARIMA model.
4. Use the forecasting insights to align inventory levels with anticipated sales.
5. Investigate any outliers or anomalies observed in the data, especially in the year 2013. Understanding the factors contributing to exceptionally high or low sales in specific months can provide valuable insights for future planning.

Next Steps

1. Continuously monitor the performance of the Seasonal ARIMA model as new data becomes available. Regularly assess the accuracy of the forecasts against actual sales and be vigilant for any changes in data patterns.
2. Refine and optimize the Seasonal ARIMA model if necessary. Consider adjusting model parameters or exploring other time series models if there are changes in the underlying patterns of shampoo sales.
3. Establish a feedback loop with key stakeholders, such as marketing and sales teams, to gather insights and qualitative information that may not be captured by the model.

Thank you