Forecasting Sales

The Business Statement

The challenge is to prevent common supply chain issues like overestimating demand leading to bloated inventory and high costs, or underestimating demand leading to underserved customers and sales loss.

The goal is to leverage forecasted monthly sales to synchronize supply with demand, aid in decision making to help build a competitive infrastructure, and measure company performance. We are forecasting 4 months of sales.

Plan Analysis

Because the task is to predict future values based on monthly sales from 01/2008 to 09/2013, the business problem was evaluated for time series. It is a candidate based on the following criteria:

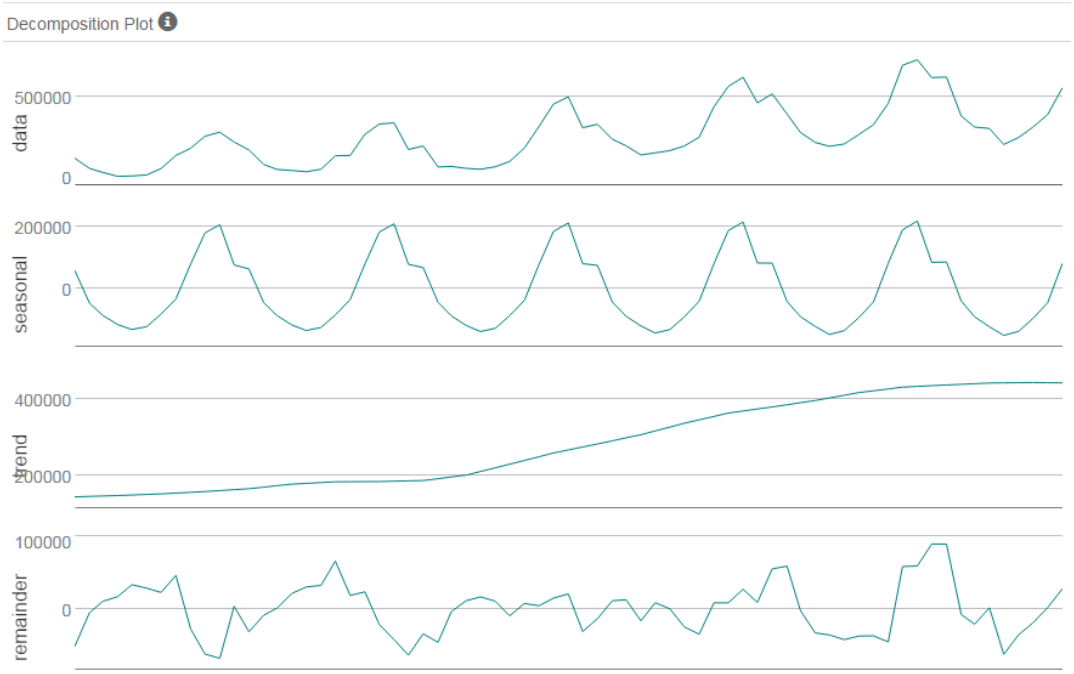
- It was over a continuous time interval
- There are sequential measurements across the interval
- There is equal spacing between every two consecutive measurements
- Each unit within the time interval has at most one data point

I evaluated ETS and ARIMA time series models with a holdout sample of 4 months in order to predict 4 months of future sales.

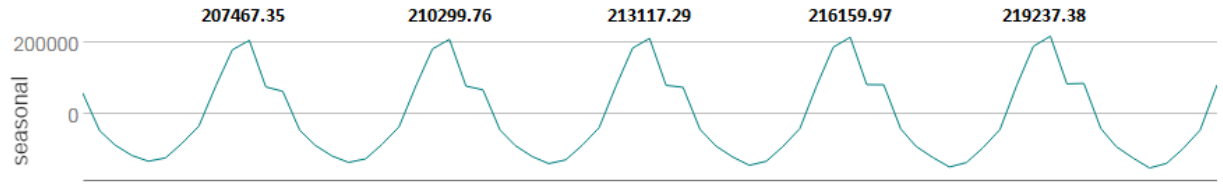
Trend, Seasonal, and Error Components

In order to accurately set the terms for the time series models, as a first step, I evaluated the data's decomposition plot by examining the trend, seasonal, and error components.

For setting the ETS model terms, the decomposition plots are the only ones needed. The ARIMA model also requires ACF and PACF plots, outlined in the ARIMA terms paragraph.



The data has a seasonal component, demonstrated in the pattern below. There is a regularly occurring spike in sales each year. It is difficult to see from the plot alone, however there are slight changes in magnitude over time. Data values were added to show the increase.



Trend shows the general course and tendency of the time series. It is the centered moving average of the time series. In this dataset, it is confirmed by the plot as upward linear trending.

The remainder plot shows the residual error that seasonal and trend plots cannot explain. It shows the changing variance or the fluctuations between large and smaller errors as the time series moves. The fluctuations for this data set are not consistent in magnitude.

Time Series Model Terms, Building the Models

ETS (M, A, M)

I selected ETS model terms (M, A, M) for the following reasons:

- The seasonal plot shows regular spikes in sales of increasing value, suggesting multiplicatively.
- The trend plot is upward linear trending for additive treatment.
- The remainder plot is not consistent in magnitude for multiplicatively.

The information criteria:

AIC	AICc	BIC
1634.6435	1645.9768	1669.4337

The in-sample errors:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056	0.1436491

ARIMA (0, 1, 1) (0, 1, 0) [12]
I selected ARIMA model terms (0,1,1)(0,1,0)[12].

The information criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

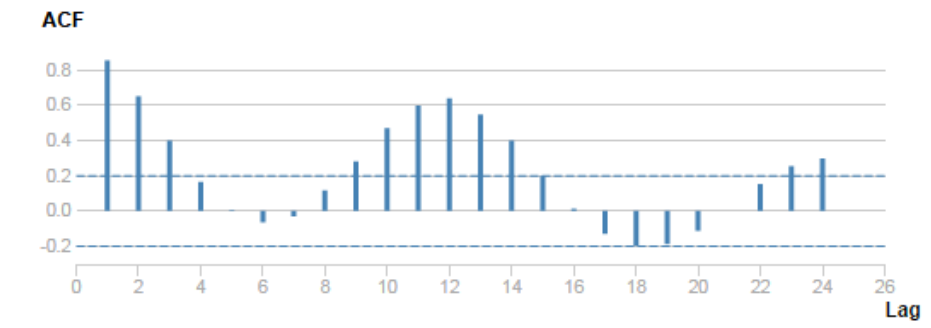
The in-sample errors:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

The following steps were taken to arrive at those model terms.

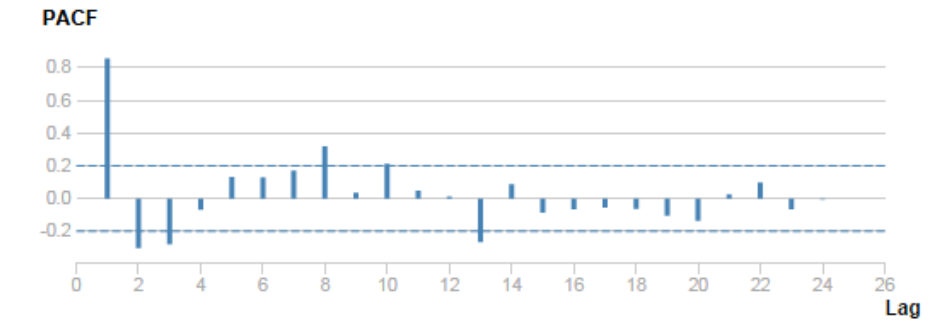
We observed seasonality in the data within the seasonal decomposition plots earlier. This can also be seen in the ACF plots without differencing, because the upward peaks at lag 1, 12, and 24 which are greater than 0.2.

Autocorrelation Function Plot



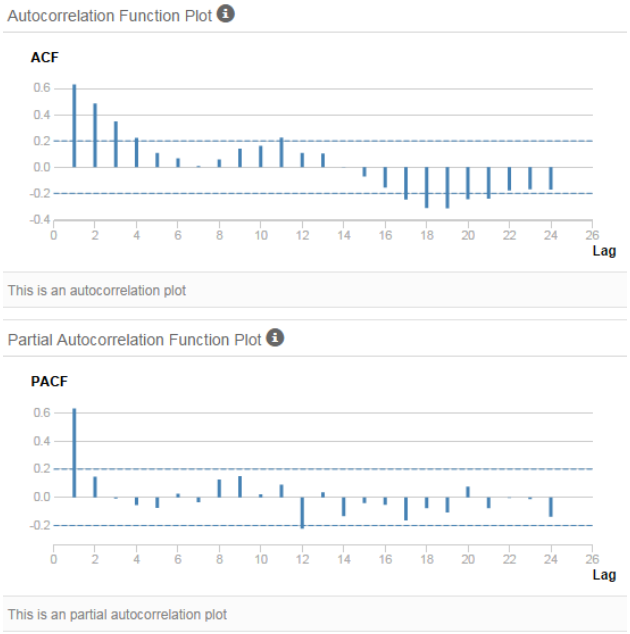
This is an autocorrelation plot

Partial Autocorrelation Function Plot



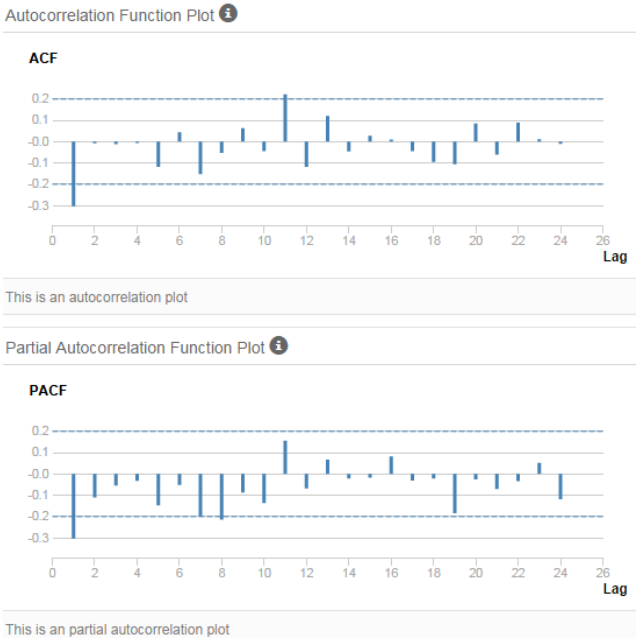
This is an partial autocorrelation plot

After taking the first seasonal difference, lag 12 and 24 no longer show seasonality and are within the acceptable range. However, because there are many points outside of .2 and -.2, a non-seasonal difference is needed for the data to become stationary.



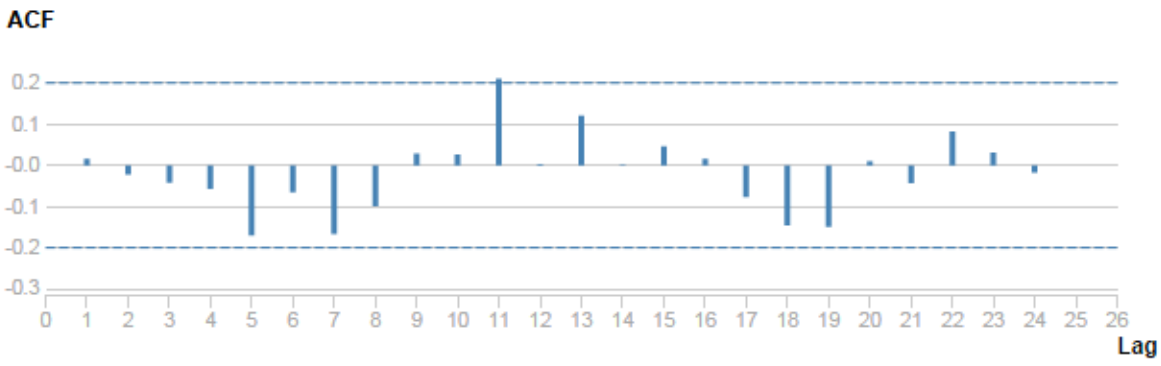
After performing a non-seasonal difference, our data is now stationary. There is a small peak at lag 11 showing that there is some residual correlation in our data, but this is acceptable and doesn't require another difference.

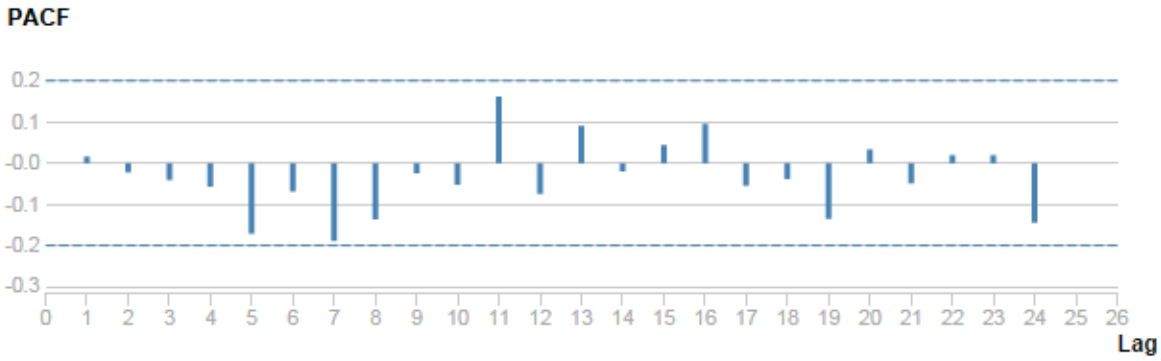
With a negative value at lag 1, a quick drop to 0 in the ACF and a more gradual trend to 0 in the PACF, the data displays MA characteristics.



The final ACF and PACF plots for ARIMA (0, 1, 1) (0, 1, 0) [12] illustrates we have the best model terms. The data is stationary. There is no longer a seasonal component and there is no significantly correlated lags. We do not need to add additional AR() or MA() terms.

Autocorrelation Function Plot





Forecast

Without the holdout sample, I would have thought the ETS (M, A, M) model was the best. It had a lower RMSE and MASE value.

ETS (M, A, M) in-sample errors without holdout sample

ME	RMSE	MAE	MPE	MAPE	MASE
3729.2947922	32883.8331471	24917.2814212	-0.9481496	10.2264109	0.3635056

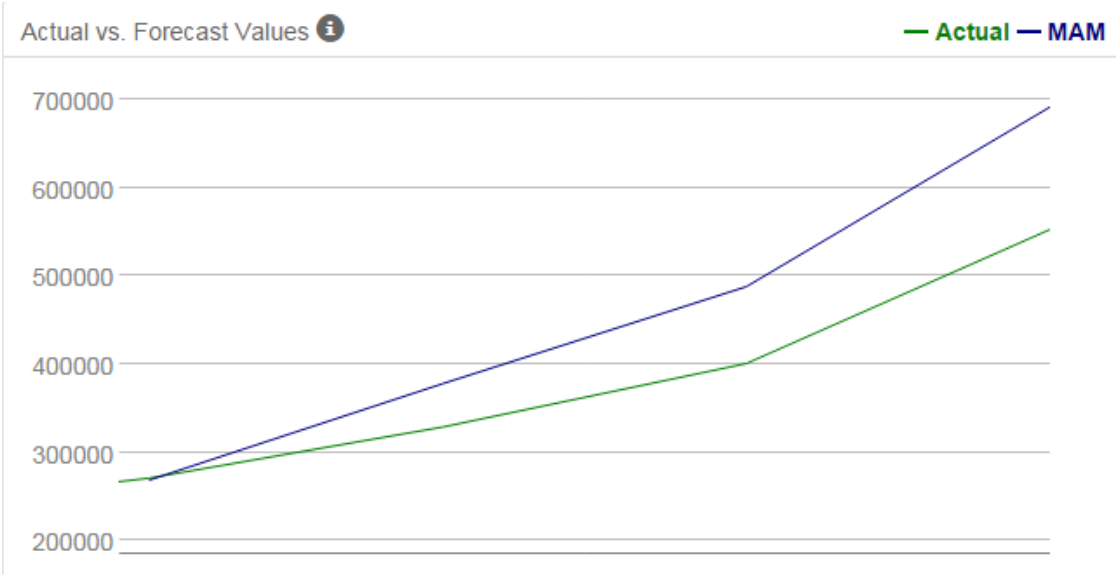
ARIMA (0, 1, 1) (0, 1, 0) [12] in-sample errors without holdout sample

ME	RMSE	MAE	MPE	MAPE	MASE
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109

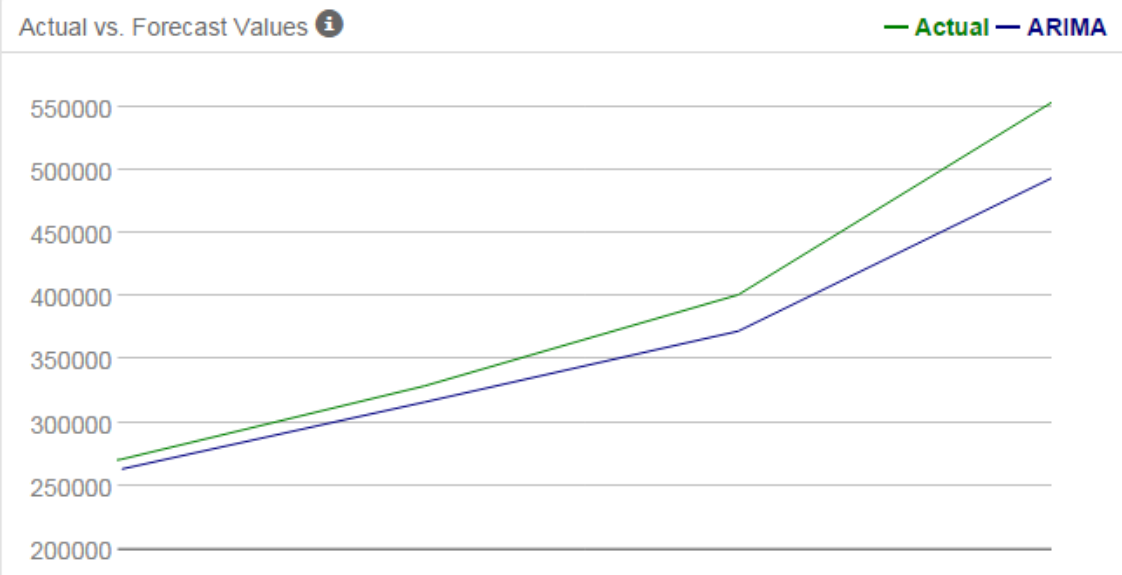
However, after testing the ETS (M, A, M) and ARIMA (0, 1, 1) (0, 1, 0) [12] models against the holdout sample, the latter performed the best. It had the smallest errors overall in the accuracy measures and the MASE was acceptable.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
MAM	-68257.4703	85623.175	69392.7195	-15.2446	15.6635	1.1532
ARIMA	27271.5199	33999.7911	27271.5199	6.1833	6.1833	0.4532

The ETS model overestimated sales values. The negative Mean Error, negative Mean Percent Error, and graphs below illustrate this bias. The MASE was also 1.1532, ideally this value should be < 1.



The ARIMA model underestimates sales. However the RMSE is much less at 33999.79 versus 85623.18, indicating that the standard deviation of the differences between predicted and actual values are much better in the ARIMA model.



Using 95% and 80% confidence intervals, the forecast for the next four periods is as follows:

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
6	10	754854.460048	833335.856133	806170.686679	703538.233418	676373.063963
6	11	785854.460048	878538.837645	846457.517118	725251.402978	693170.082452
6	12	684854.460048	789837.592834	753499.24089	616209.679206	579871.327263
7	1	687854.460048	803839.469806	763692.981576	612015.938521	571869.450291