Project: Forecasting Sales

Step 1: Preparation

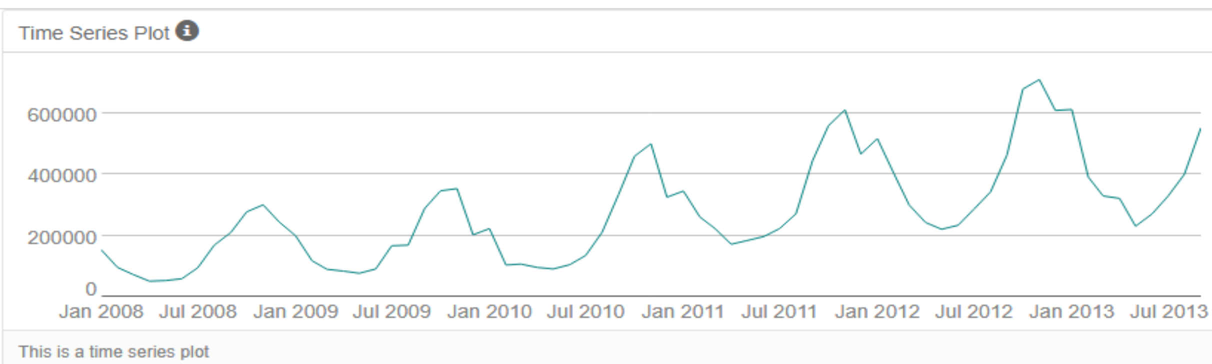
In the first step I checked the data to evaluate whether it is appropriate to use time series models. Then I split the data for validation purposes.

The data I have in this project seem to be in the form of a continuous time series data as the intervals between each value are fixed and equal and there is no missing value within the sequence.

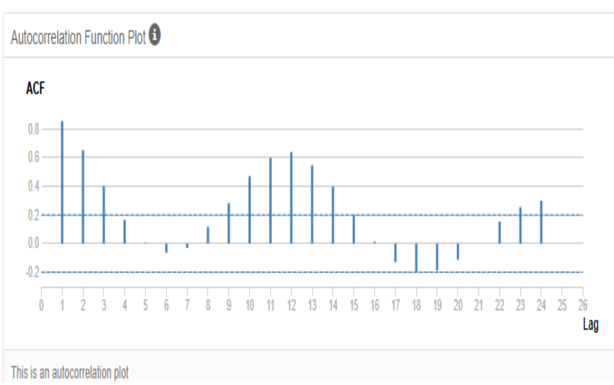
I used the last four cases as holdout sample in this case as we predict four periods. Usually the hold out sample/validation sample is the same length of period with the period that will be forecast.

Step 2: Determining Trend, Seasonal, and Error components

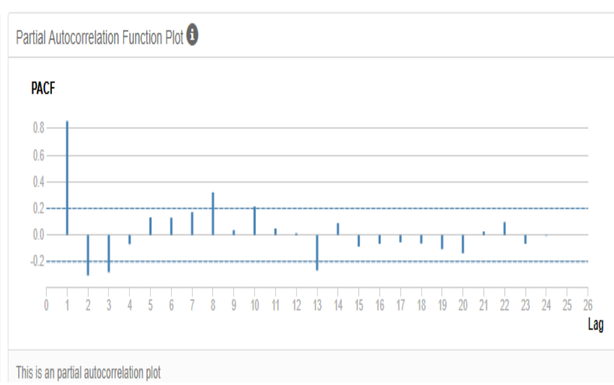
I graphed the data set to decompose the time series into its three main components: trend, seasonality, and error.



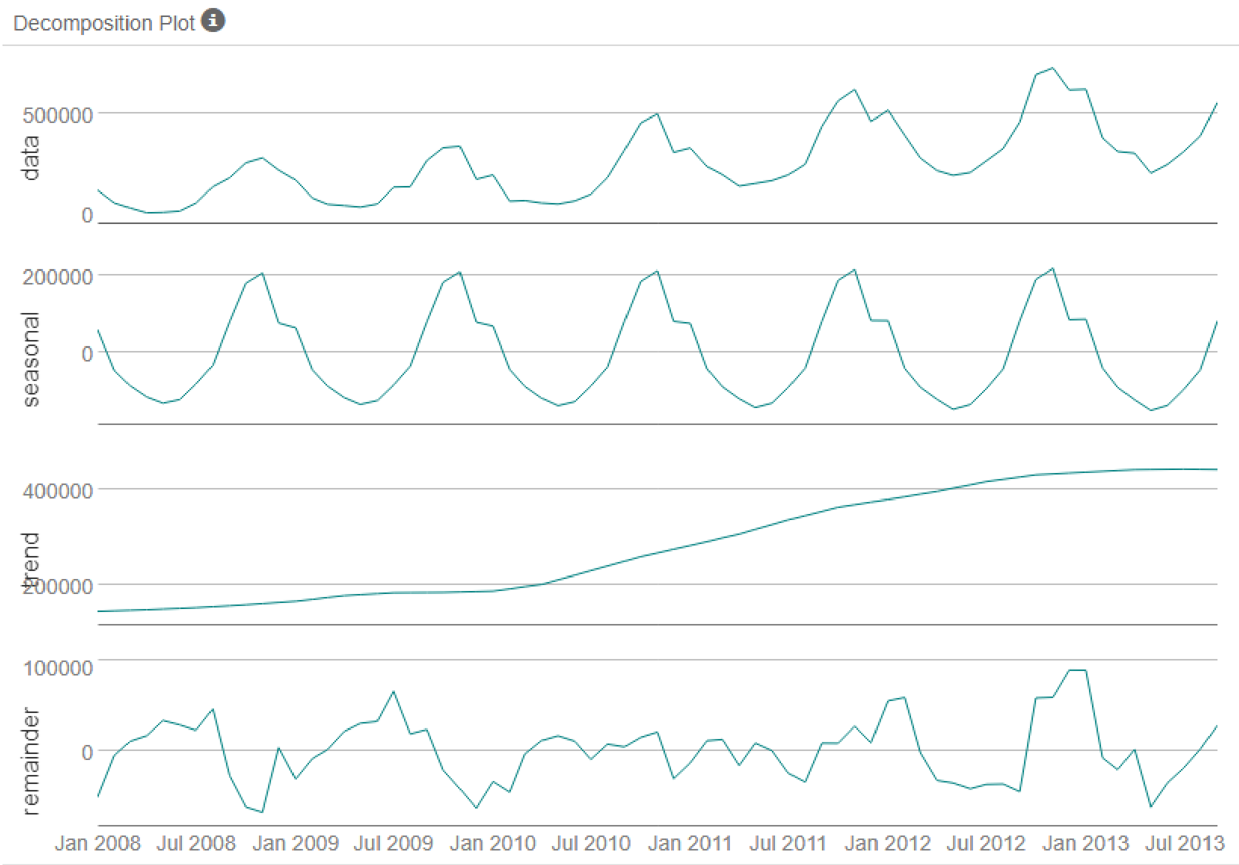
This is a time series plot



This is an autocorrelation plot



This is an partial autocorrelation plot



This is a decomposition plot

As can be seen from the data, there is an upward trend in the time series. There is also seasonality as the series peak in November each year. The decomposition plot also confirms our initial analysis indicating both trend and seasonality.

Since our series have seasonality, we are supposed to apply seasonal differencing for the ARIMA models. When we look at the data, we see a change in magnitude over time suggesting the use of multiplicative method for the ETS models. The error plot also shows variations in the error in magnitude suggesting the use of multiplicative method for the ETS models.

Step 3: Building the Models

After analyzing the graphs, I determined the appropriate measurements to apply to ARIMA and ETS models and described the errors for both models.

The decomposition plots indicate a linear upward trend suggesting an additive method for the ETS model. As for seasonality we see a variation in magnitude suggesting a multiplicative method. IN the same way the variation in the error terms in magnitude also suggest a multiplicative method. Therefore, the best model seems to be ETS (M, A, M) model.

RMSE is 40906.5798 indicating a variance which is about 41000 units around the mean.
MASE on the other hand is 0.41 indicating a strong model to forecast which is below 1.00 which is the upper threshold used for model accuracy.

Method:
ETS(M,A,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
281.3033372	40906.5798105	28212.6317577	-0.9199379	11.2444007	0.4115798	0.2971592

Information criteria:

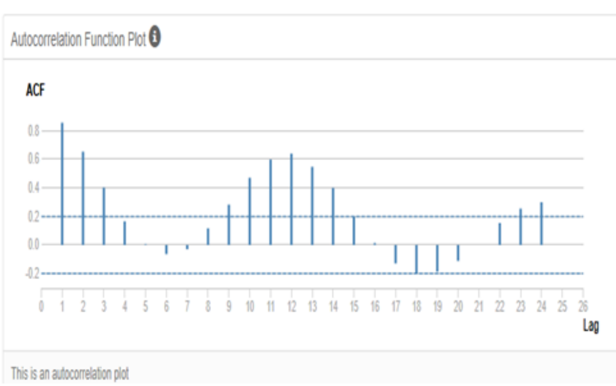
AIC	AICc	BIC
1649.9458	1662.9671	1686.9104

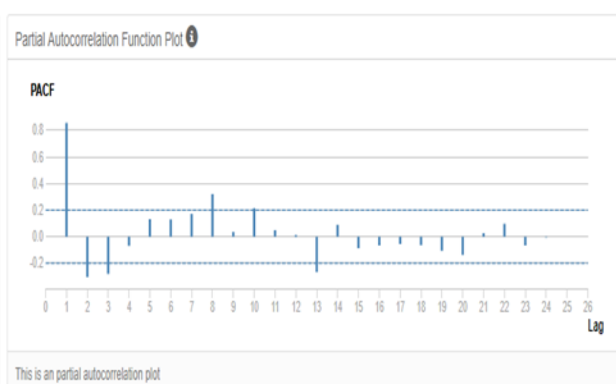
I used Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component to justify the model terms I chose.

Our series has seasonality components therefore I will use ARIMA (p, d, q) (P, D, Q)S model to forecast.

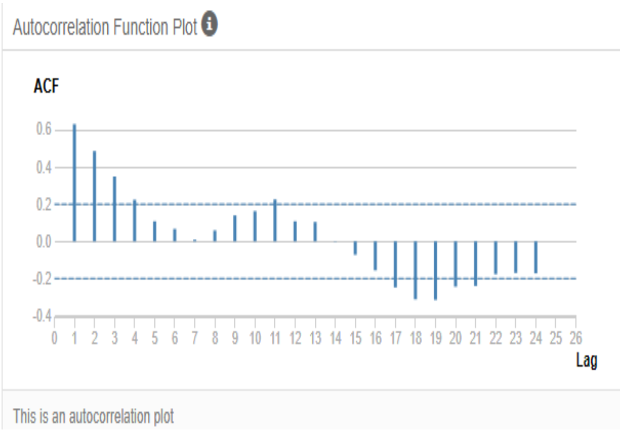
Step 4: Forecasting

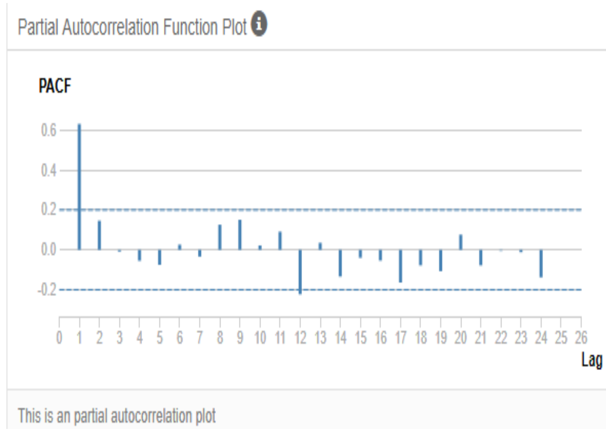
I compared the in-sample error measurements to both models and compare error measurements for the holdout sample in the forecast. I chose the best fitting model to forecast the next four periods.
The ACF plot decays slowly towards Zero and there seems to be autocorrelation among the seasonal lags. Therefore, we need to apply differencing to obtain a stationary series.



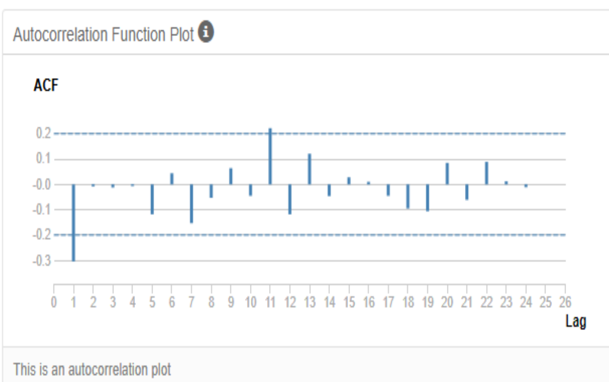


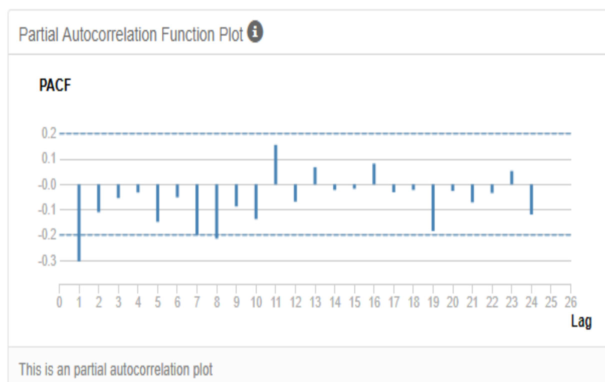
After the seasonal differencing we do not see so much difference in the ACF and PACF plots. There still seems to be autocorrelation. Therefore, I applied another differencing.





After the seasonal first differencing ACF and PACF plots do not indicate autocorrelation problem. Most of the significant legs have been removed. It seems that we do not need any further differencing.





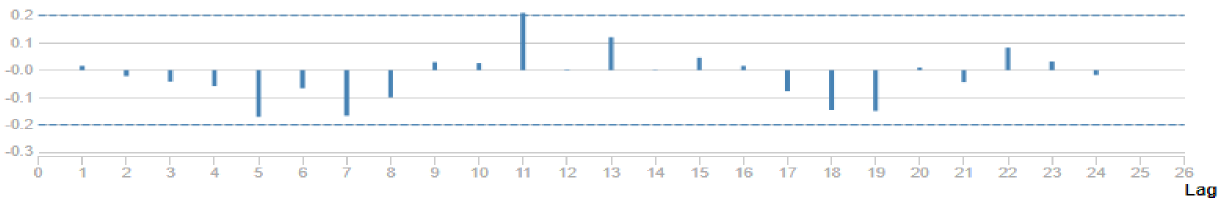
Now looking at these plot we can determine the ARIMA terms we will add to the model. Accordingly, since both ACF and PACF plots indicate a negative significant correlation at lag 1, and since there is one significant lags when both plots are considered, this requires MA (1) model. The seasonal lags are not significant therefore we will not need to add any seasonal autoregressive or moving average terms.

In the final analysis the proper ARIMA model seems to be ARIMA(0, 1, 1)(0, 1,0)12

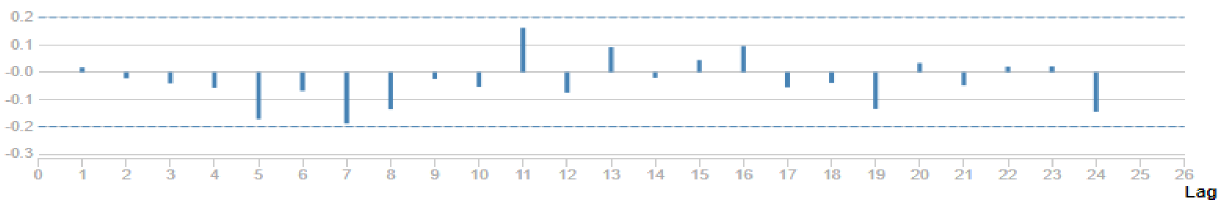
The ACF and PACF plots of the ARIMA models suggest no additional AR or MA terms as there is no autocorrelation among the lags.

Autocorrelation Function Plot

Autocorrelation Function



Partial Autocorrelation Function



Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

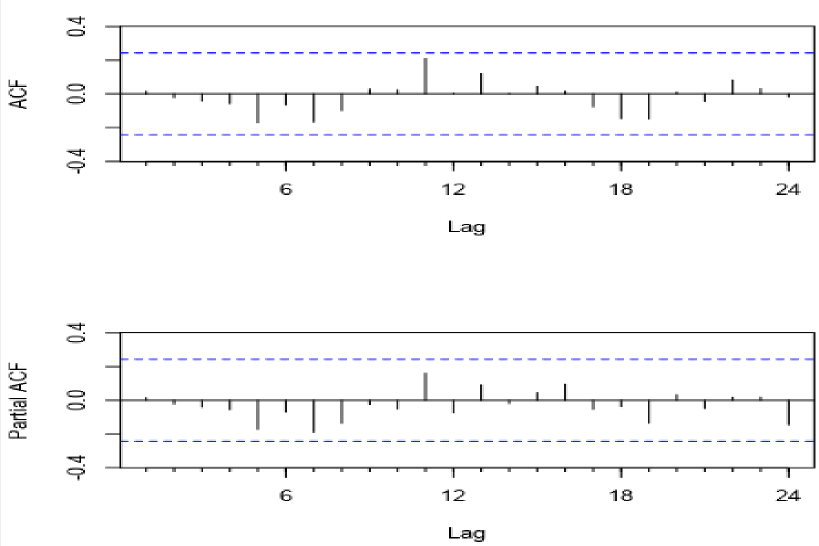
In-sample error measures:

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

Ljung-Box test of the model residuals:
Chi-squared = 16.4458, df = 23, p-value = 0.83553

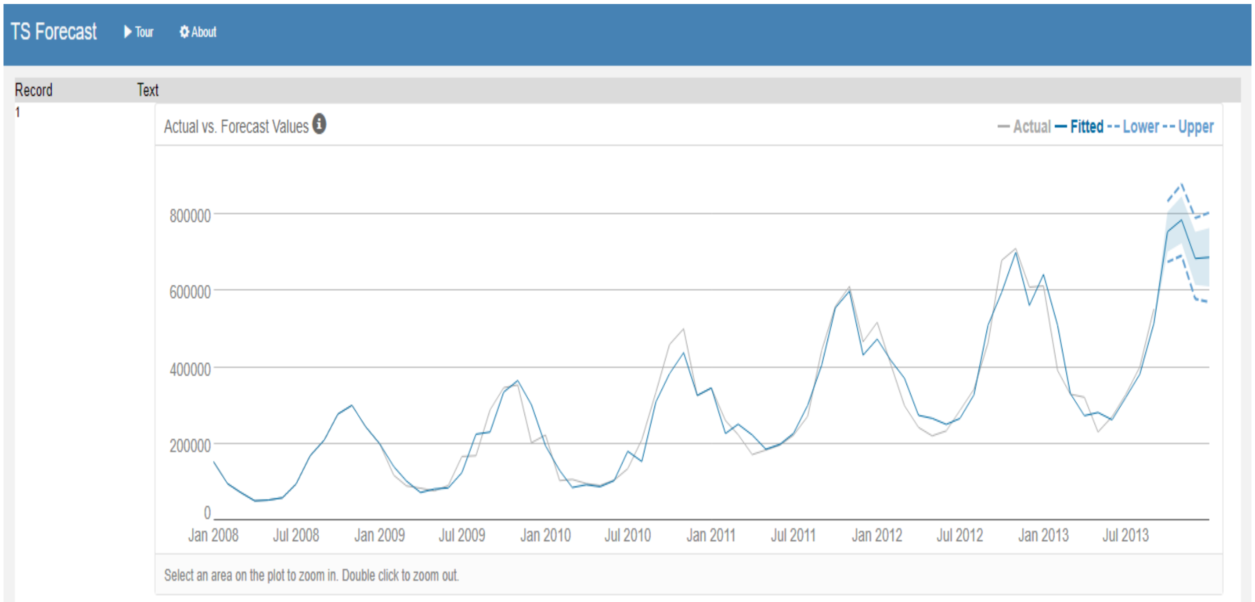
Plots

Autocorrelation Function Plots



In the ARIMA models when we check the RMSE and MASE which can be used to compare forecasts of different models we see that variance is 36761.528.

The MASE is 0.36 and is significantly lower than 1.00 which is a commonly used threshold for model accuracy, indicating a stronger model than ETS.



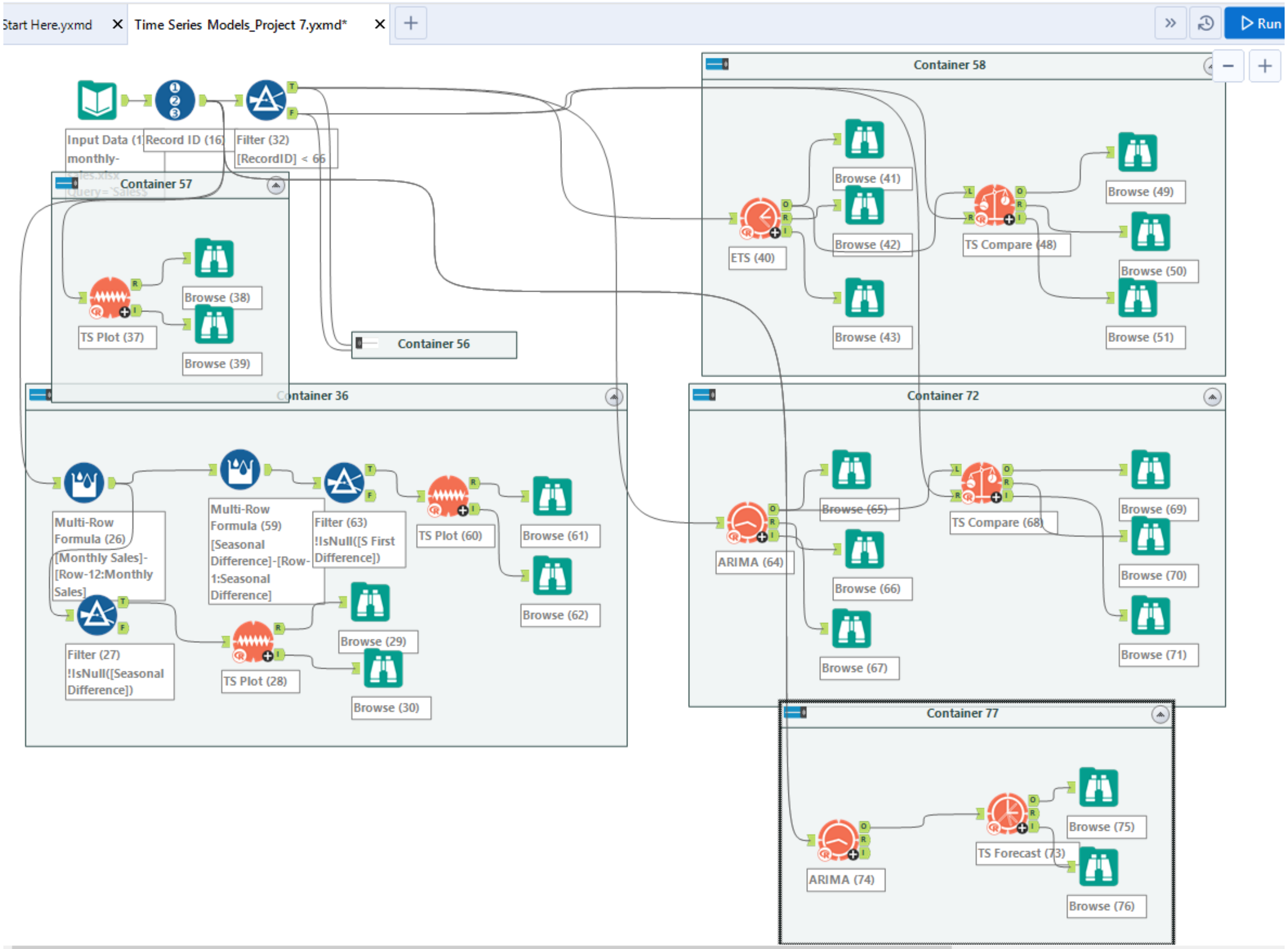
Forecast results using 95% and 80% confidence intervals are shown below.

7 of 7 Fields · Cell Viewer · 4 records displayed, 1742 bytes

Record	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2013	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
2	2013	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
3	2013	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
4	2014	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462

The display value was rounded to 6 decimal places for clarity.

Below is my Alteryx workflow indicating the analytical process.



PREPARED BY **KADIR AKYUZ**