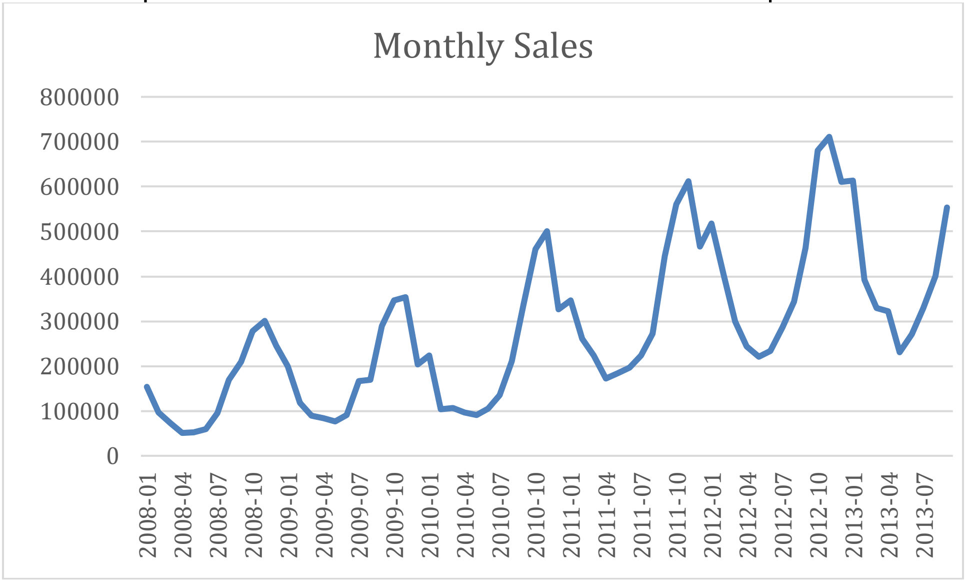
Project: Forecasting Sales

Step 1: Plan Your Analysis

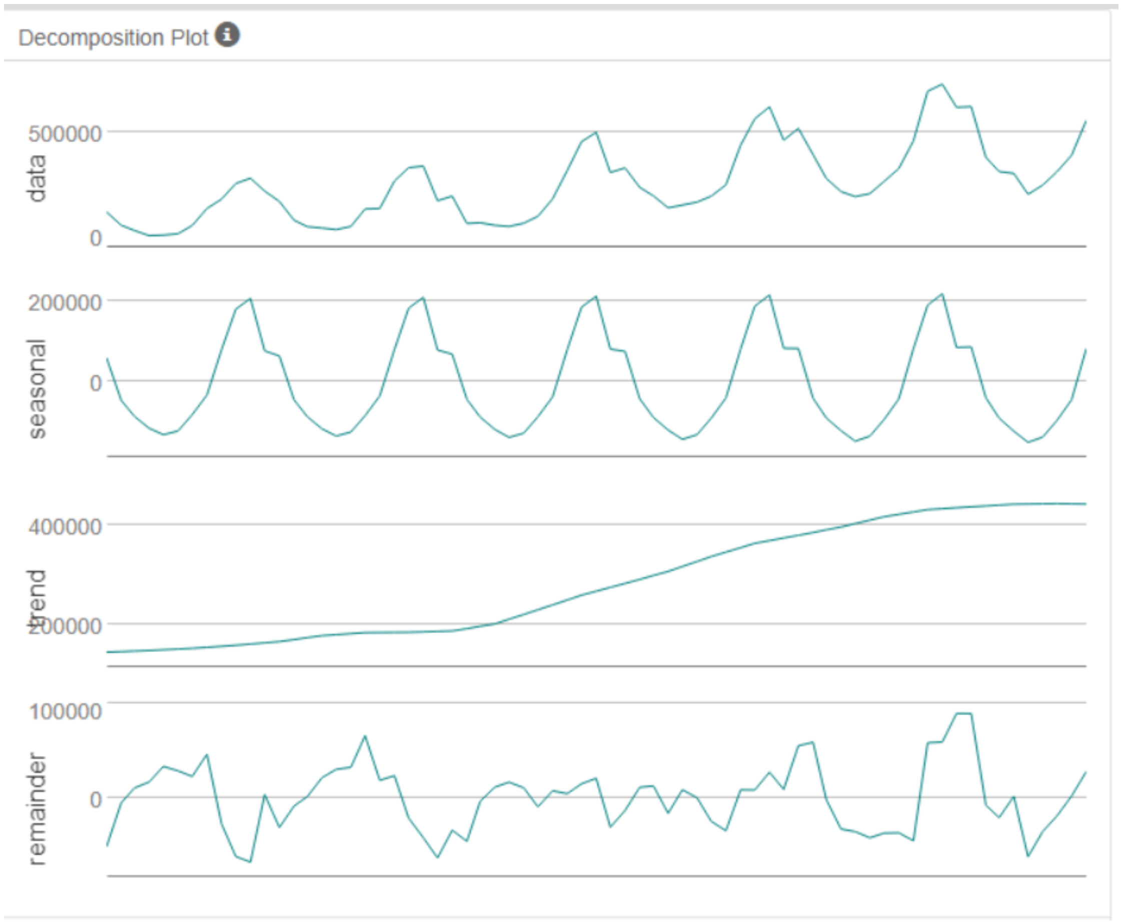
We have a file that contains monthly sales data of a store from January 2008 to September 2013. If we have a look at the data, we will see that the data is ordered over time, is sequential and has equal intervals. Let us have a look at visualization plot:



As we see, there is a clear uptrend which means that if we draw a line through the heights of the graph we will have a general direction going up. Also the data has a seasonal pattern: if we notice its heights are all around month October and lower parts are all around April, which means that the same trend repeats over time. In this data we don’t have any cyclical or random variations. We can conclude that our data meets the criteria of time series dataset and can be used in time series models.
As we are asked to forecast data for the next four months, so as a holdout sample we will use the last four months’ data: from 2013/06 to 2013/09.

Step 2: Determine Trend, Seasonal, and Error components

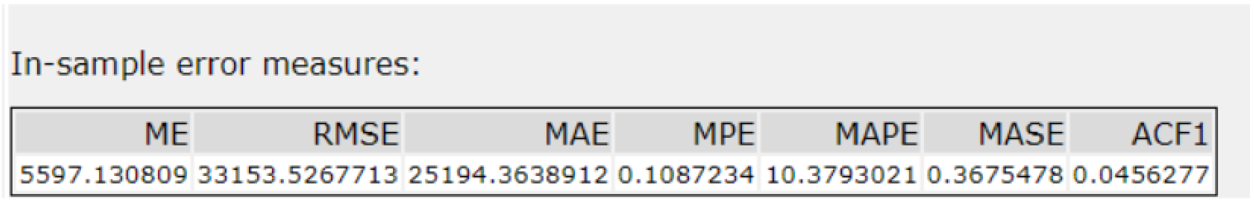
Using the TS Plot tool in alteryx we can decompose the time series into its main components: trend, seasonality, and error. Let us have a look at decomposition plot:



As we see, there is a strong seasonality in our dataset, and if we look closer we will see that it has a multiplicative behavior, as the values (heights) are slightly increasing over time. It is obvious, as well, that there is an upward trend in our dataset, and we can notice that in the recent three years its increase is greater than in the first years. Also the trend component has an additive behavior. There is no any pattern in error plot, the values are randomly distributed around 0 which means that the mean is close to 0, and all together we can conclude that there should not be obvious errors between forecasted and actual values. Also we can see from the plot that the error has multiplicative behavior.

Step 3: Build your Models

First we will build ETS model using our dataset in Alteryx. As we have discussed above and based on the decomposition plot we see that the Error component is multiplicative, the Trend component is additive and the Seasonal component is multiplicative, so our ETS model terms will be ETS (M, A, M), so we will manually choose the model type. After running the model, we can have a look at in-sample errors:

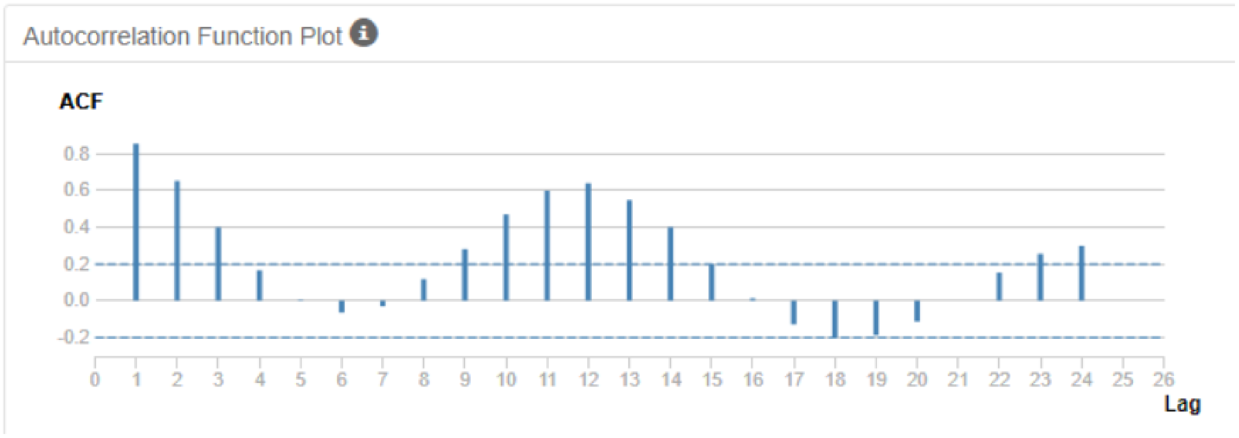


In-sample error measures:

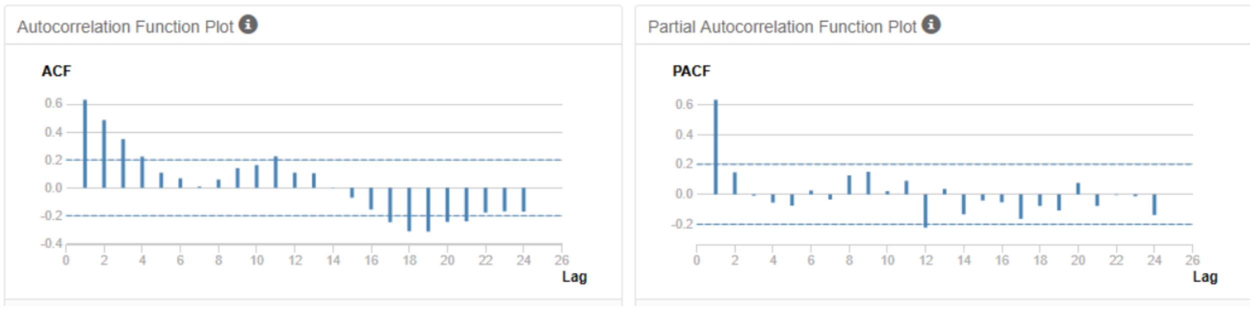
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

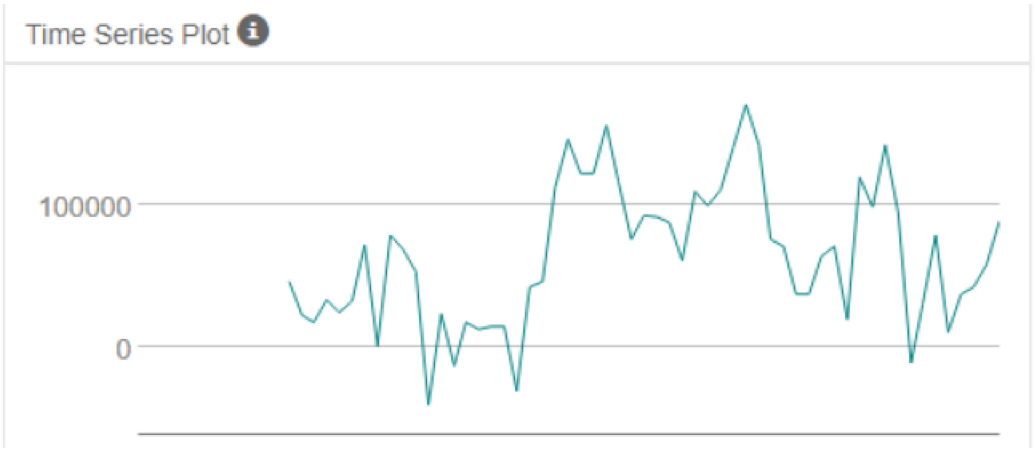
The Root Mean Squared Error (RMSE) of the model is 33153.5267713, and the Mean Absolut Scaled Error (MASE) is 0.3675478. AIC value is 1639.5.

Now let us build an ARIMA model from the same data. To decide on the model terms, we first need to look at ACF and PACF graphs.

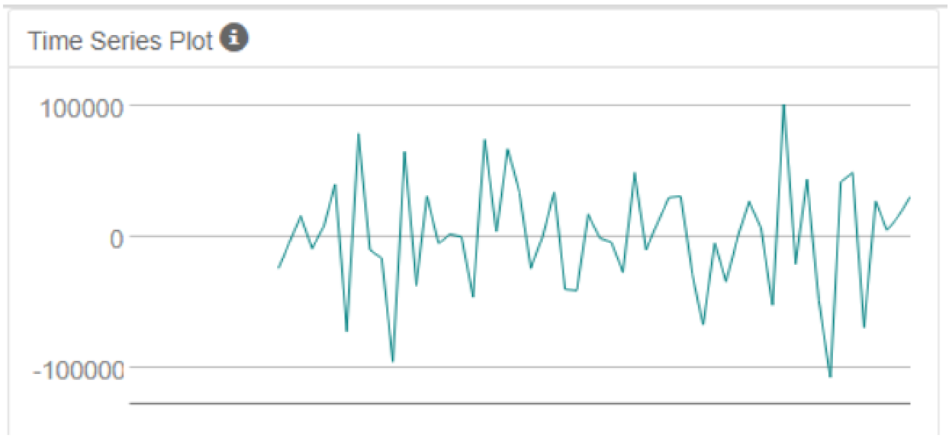


As we can see there is an autocorrelation between the first few lags, and if we notice, it repeats in seasonal lags. So we need to do differencing to make the data more stationary. We will start with seasonal differencing:

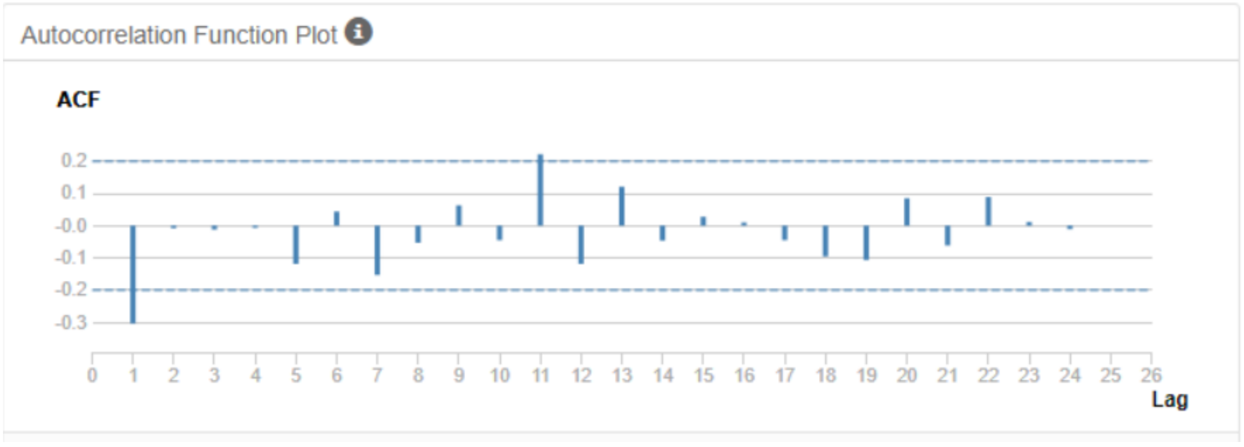


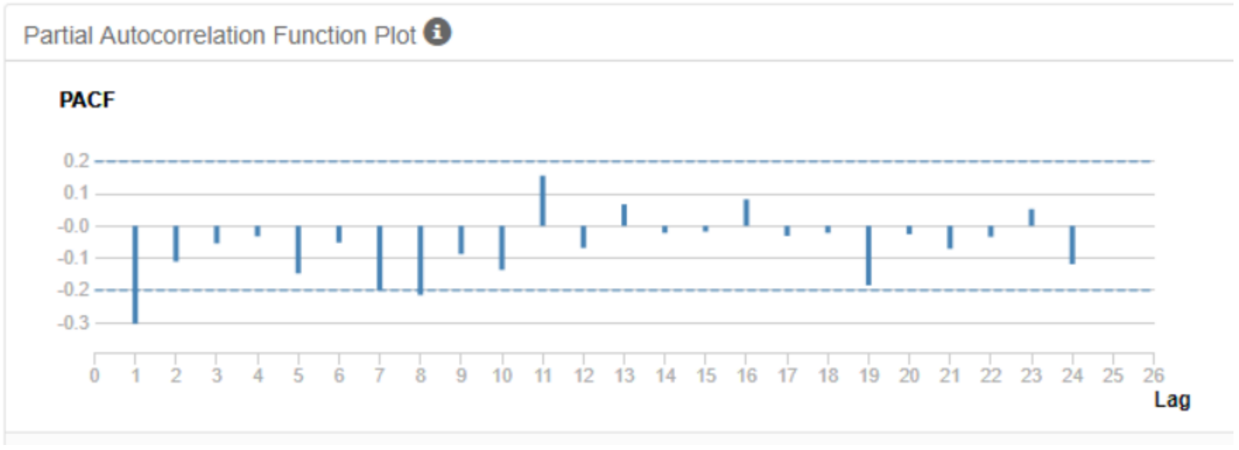


As we can see, the seasonal differencing decreased the level of autocorrelation, however, it still exists. We can prove that looking at the time series plot: the data still is not stationary. This means that we need to do another differencing, this time non_seasonal.
Checking the time series plot, we see that the dataset is sufficiently stationary and we do not need to do further differencing.

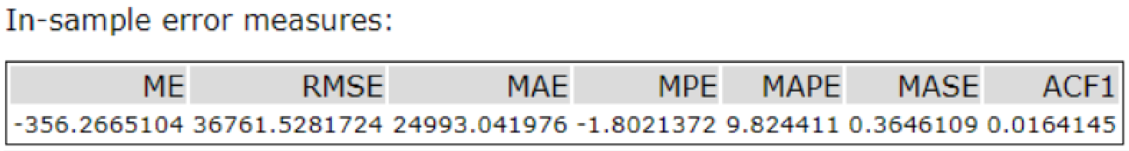


Now let us have a look at ACF and PACF graphs after first differencing of seasonal differencing and decide the terms of time series and seasonal model terms:





Looking at ACF graph we see a significant negative correlation at lag 1 and if we look at PACF we see that it slowly decays, so here we have MA term. As in ACF graph the autocorrelation drops off after first lag, we have MA (1). We did one non-seasonal differencing, so I term will be one. Thus, the non-seasonal component of ARIMA model or (pdq) values are ARIMA (0,1,1).
Looking at Lags 12, 24 in both ACF and PACF graphs we don't see any significant correlation neither positive nor negative, this means that the terms, or (PDQ) are (0,1,0) as we did one seasonal differencing. So our final model components are ARIMA (0,1,1) (0,1,0) [12]
Now, we can build the ARIMA model in Alteryx and look at the in-sample errors:

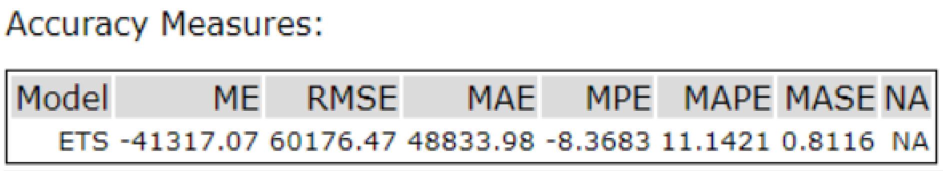

In-sample error measures:

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

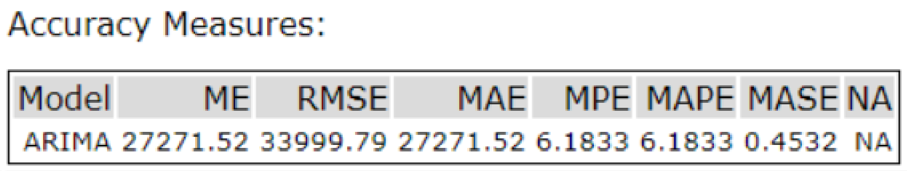
The Root Mean Squared Error (RMSE) of the model is 36761.5281724, and the Mean Absolut Scaled Error (MASE) is 0.3646109. AIC value is 1256.6.

Step 4: Forecast

Comparing the in-sample errors of both models, particularly RMSE and MASE errors, as well as AIC values we see that both models have almost the same MASE errors, around 0.36. This is a good indicator, as it is lower than 1. RMSE errors are close, however, ARIMA models RMSE error is higher, around 36761 compared to ETS model's 33153. And if we compare AIC values the one for ARIMA is lower 1256 a better indicator than in case of ETS model's 1639.
Now we can use TS-Compare too in Alteryx to validate each model against the 4 holdout periods:

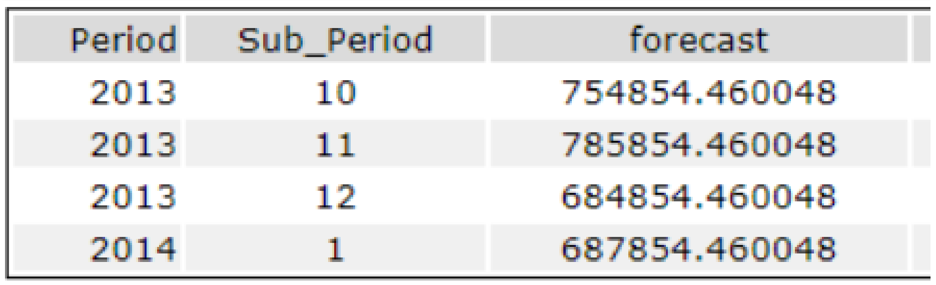

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA


Accuracy Measures:

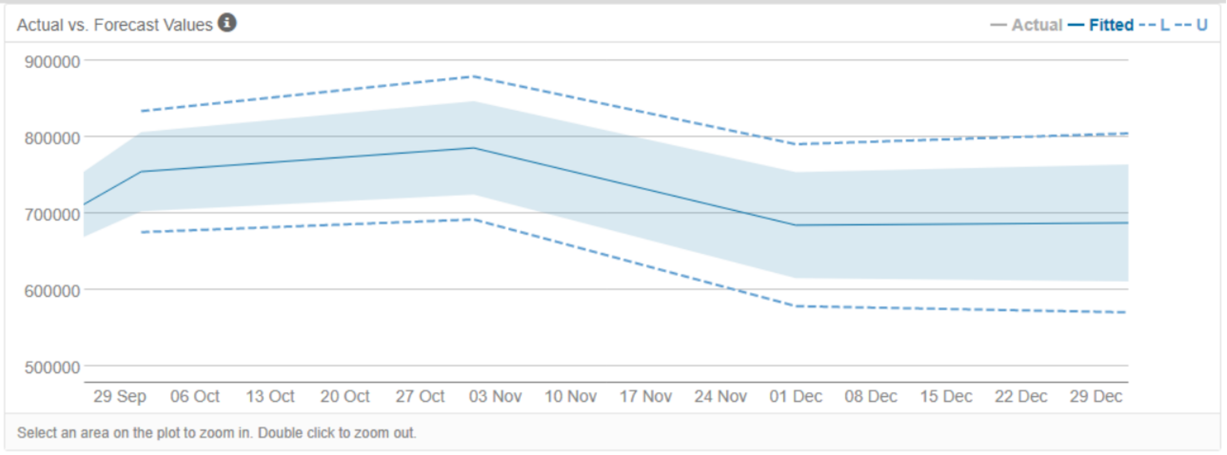
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

Comparing the accuracy measures for ETS and ARIMA models for holdout samples, we see that both RMSE and MASE errors have much better figures in case of ARIMA model: 33999.79 and 0.4532. So, comparing both in-sample errors, accuracy measures of TS-compare and AIAC values, we will choose ARIMA model for our forecasting.
Using ARIMA model and TS Forecast tool in Alteryx we can forecast the values for 4 months (October/2013 to January /2014).

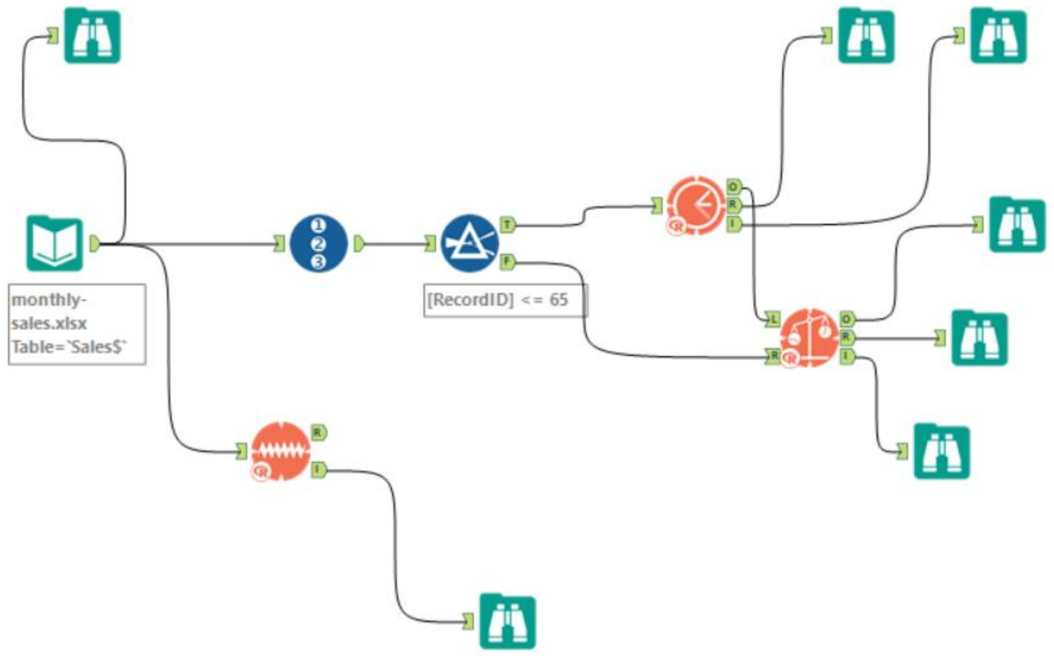


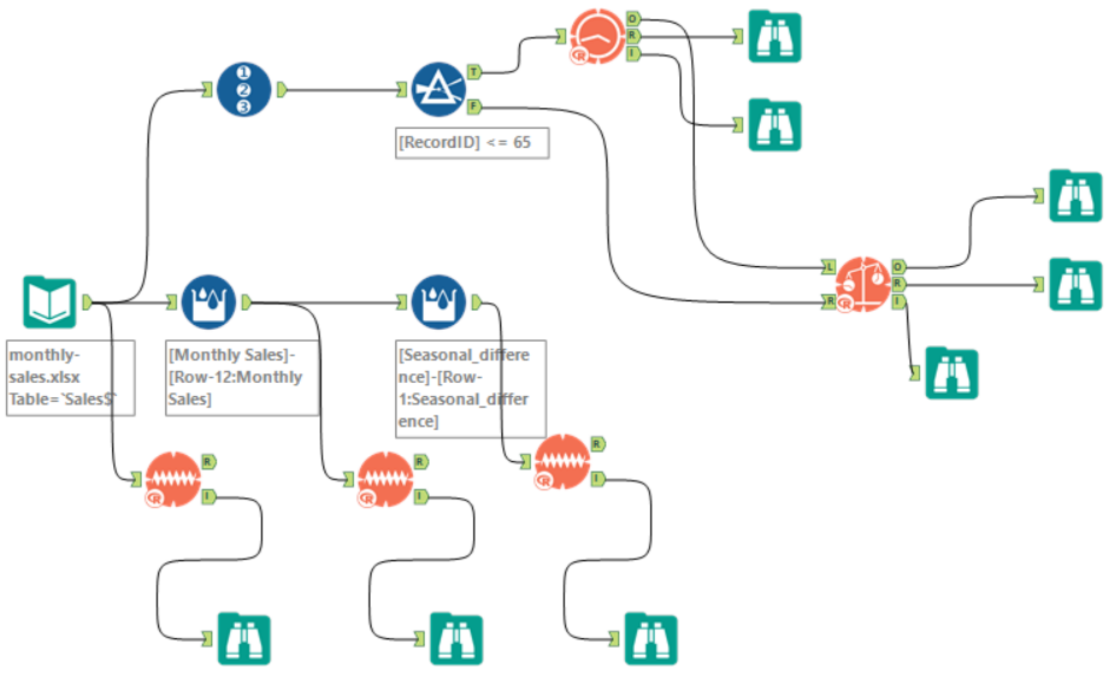
Period	Sub_Period	forecast
2013	10	754854.460048
2013	11	785854.460048
2013	12	684854.460048
2014	1	687854.460048

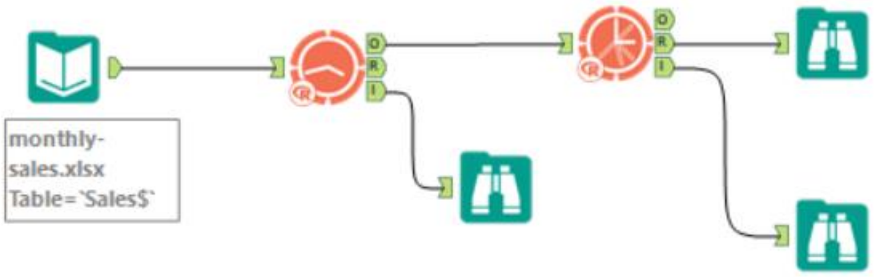
And here is the graph representing 80% and 95% confidence intervals:



Below are Alteryx workflows used in this project:


monthly-sales.xlsx Table=`Sales$`
[RecordID] <= 65


monthly-sales.xlsx Table=`Sales$`
[Monthly Sales]-[Row-12:Monthly Sales]
[Seasonal_difference]-[Row-1:Seasonal_difference]
[RecordID] <= 65


monthly-
sales.xlsx
Table=`Sales$`
O
R
I
O
R
I